Process Control

Jean-Pierre Corriou

Springer

Process Control

Springer-Verlag London Ltd.

Jean-Pierre Corriou

Process Control

Theory and Applications

With 339 Figures

12 3

Prof. Dr. Jean-Pierre Corriou
Laboratoire des Sciences du Génie Chimique
Ecole Nationale Supérieure des Industries Chimiques
LSGC-CNRS-ENSIC
1, rue Grandville BP 451
54001 Nancy Cedex
France

British Library Cataloguing in Publication Data
Corriou, Jean-Pierre
Process Control : theory and applications
1. Chemical process control
I. Title
660.2'815
ISBN 978-1-84996-911-6 ISBN 978-1-4471-3848-8 (eBook)
DOI 10.1007/978-1-4471-3848-8

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

ISBN 978-1-84996-911-6

springeronline.com

© Springer-Verlag London 2004
Originally published by Springer-Verlag London Berlin Heidelberg in 2004

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typesetting: Electronic text files prepared by author

69/3830-543210 Printed on acid-free paper SPIN 10950364

Preface

Organization of the Book

This book has been conceived to progressively introduce concepts of increasing difficulty and allow learning of theories and control methods in a way which is not too brutal. It contains different levels of reading (Figure 1). In particular, the majority of the first part can be undertaken by students beginning in control or by technicians and engineers coming from the industrial world, having up to that point only practical contact with control and a desire to improve their knowledge. The subsequent parts need a minimum previous knowledge in control. They also allow the use of often higher-performance techniques. Without pretending to be exhaustive, this book proposes a wide range of identification and control methods applicable to processes and is accompanied by similar typical examples that provide comparison elements.

This book does not pretend to compete with theoretical control books specialized in one subject or another, e.g. identification, signal processing, multivariable control, robust control or nonlinear control. However, the readers whoever they are, undergraduate or graduate students, engineers, researchers, professors, will find numerous references and statements that enable understanding and application in their own domain of a large number of these concepts, taking their inspiration from the cases treated in the present book.

Several controls are examined under different angles:

- single-input single-output internal model control in continuous and discrete time, multivariable internal model control in discrete time,
- pole-placement control in continuous and discrete time,
- single-input single-output linear quadratic control by continuous transfer function and multivariable state-space linear quadratic control in continuous and discrete time,
- single-input single-output generalized predictive control, multivariable model predictive control, possibly with observer, linear or nonlinear.

The consideration of the same problem by different approaches thus enables

a thorough examination.

Examples are most often taken from the chemical engineering domain, and include in general chemical, biological and polymerization reactors, a catalytic cracking reactor (FCC) and distillation columns. These examples are detailed, even at the numerical level, so that the used reasoning can be verified and taken again by the reader. Simulations have been realized by means of Matlab© and Fortran77 codes.

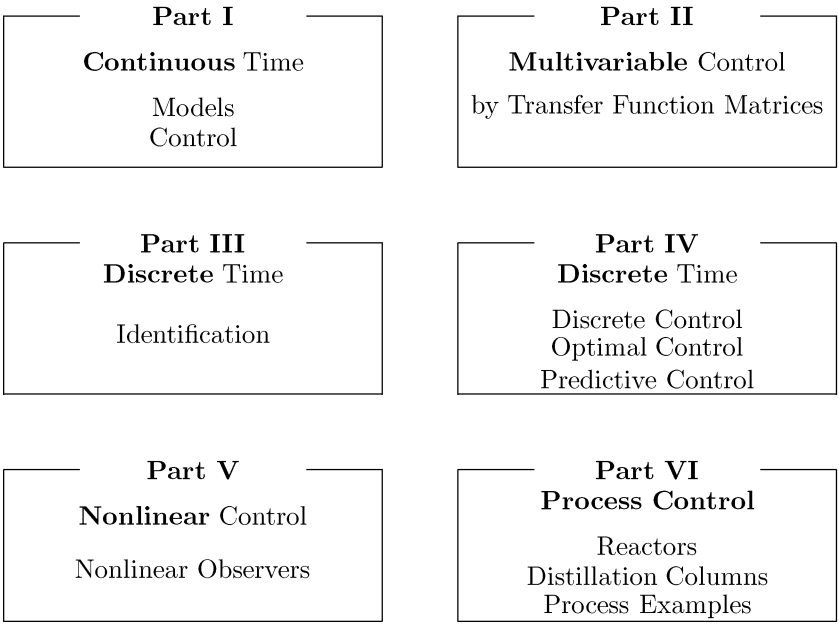


Figure 1: General organization of the book

Part I concerns single-input single-output continuous-time control. The chosen presentation of single-input single-output feedback linear process control is classical and voluntarily simple for most of the part. It presents the pedagogical advantage of decomposing the approach to a control problem, introduces a given number of important notions and, according to our opinion, facilitates the understanding of discrete-time control, called digital control, and of nonlinear control. Also, it is close, in its conception, to a large part of industrial practice, at least in the domain of chemical engineering. Continuous PID control is abundantly treated, but without exclusivity. The main types of dynamic models met in chemical engineering are commented on and system models are presented as well in state space as transfer functions (Chap. 1). Control is first simply related to the PID controller (Chap. 2). Stability is presented both for linear and nonlinear systems. Thus, the stability of a polymerization reactor is detailed by displaying the multiple stationary states in relation to the physical behaviour of the reactor (Chap. 3). The design of controllers first deals with PID, then is broadened to internal model control, which is very

Table 1: Contents of the chapters

Part I

Dynamic Modelling, Open-Loop Dynamics	Chapter 1
PID Linear Feedback Control	Chapter 2
Linear and Nonlinear Stability Analysis	Chapter 3
Design of Feedback PID Controllers, Pole-Placement	Chapter 4
Linear Quadratic Control, Internal Model Control	
Frequency Analysis, Robustness	Chapter 5
Improved Controllers, Smith Predictor	Chapter 6
Cascade, Feedforward Control	
State Representation, Controllability, Observability	Chapter 7
Realizations and Model Reduction	

Part II

Multivariable Control	Chapter 8
by Transfer Function Matrices	

Part III

Discrete-Time Generalities	Chapter 9
Signal Processing	
Identification Principles	Chapter 10
Identification Models	Chapter 11
Identification Algorithms	Chapter 12

Part IV

Discrete Pole-Placement	Chapter 13
Discrete PID	
Discrete Internal Model Control	
Optimal Control	Chapter 14
Continuous LQ and LQG Control	
Discrete LQ and LQG Control	
SISO Generalized Predictive Control	Chapter 15
MIMO Model Based Predictive Control	Chapter 16

Part V

Nonlinear Control	Chapter 17
Nonlinear Observers, Statistical Estimators	Chapter 18

Part VI

Reactors	Chapter 19
Distillation Columns	Chapter 20
Process Examples, Benchmarks	Chapter 21

important industrially, pole-placement control and linear quadratic control by means of continuous transfer functions (Chap. 4). Frequency analysis begins classically by the analysis in Bode and Nyquist representations, but is then

extended to robustness and sensitivity functions (Chap. 5). The improvements of controllers including time delay compensation, cascade control and feedforward control are reviewed with application examples in industrial processes (Chap. 6). The first part finishes with the concepts of state representation for linear systems, controllability and observability (Chap. 7). Some more difficult parts of different chapters such as robustness, pole-placement control or linear quadratic control can be tackled in a subsequent reading.

Part II consists of only one chapter which deals with multivariable control by either continuous or discrete transfer functions matrix. This choice was made because of the relatively common practice of system representation chosen in process control. The chapter essentially presents general concepts for the design of a multivariable control system. Other types of multivariable control are treated in specific chapters in other parts of the book: linear quadratic control and Gaussian linear quadratic control in Chap. 14, model predictive control in Chap. 16 and nonlinear multivariable control in Chap. 17. In fact, Part III can be studied before Part II.

Part III begins by considering signal processing whose general concepts are necessary in identification and control. Then, the general aspects of digital control and sampling are treated and discrete transfer functions are introduced (Chap. 9). The remainder of Part III is devoted to discrete-time identification. First, different model types are presented and the principles of identification are explained (Chap. 10), then different types of models are presented (Chap. 11). Lastly, the main algorithms of parametric identification are detailed with many indications on usage precautions (Chap. 12). The parametric identification of a chemical reactor is presented. Identification is treated as single-input single-output except with respect to the Kalman filter, which can be applied to multi-input multi-output systems.

In Part IV, several classical types of digital control are studied. Chapter 13 describes pole-placement, digital PID and discrete internal model control as single-input single-output control with application to the same chemical reactor. In Chap. 14, optimal control is considered in the general framework of dynamic optimization applicable to nonlinear continuous or discrete systems; it includes general methods such as variational methods, Euler, Hamilton-Jacobi, Pontryagin and dynamic programming. Linear quadratic and linear quadratic Gaussian control is presented in direct relation to optimal control both for continuous and discrete state-space system descriptions. An application of multivariable linear quadratic Gaussian control to an extractive distillation column with two inputs and two outputs is presented. Two types of predictive control are studied. Chapter 15 concerns generalized predictive control for single-input single-output systems represented by their discrete transfer function with application to the previously mentioned chemical reactor. Chapter 16 is devoted to model predictive control applicable to large multivariable systems known by transfer functions or state-space models. Furthermore, model predictive control is popular in industry as it allows constraints to be taken into account. Two multivariable applications for a catalytic cracking reactor (FCC) are shown.

Part V concerns nonlinear control presented through differential geometry

(Chap. 17) and state observers (Chap. 18). These are recent developments in control, and are potentially very powerful. To facilitate its approach, several concepts are analyzed from a linear point of view, then nonlinear control for a single-input single-output system is studied with input-state and input-output linearization. Nonlinear multivariable control is just outlined. State estimation is necessary in nonlinear control. Chapter 18 on observers does not concern the linear Kalman filter described in Part III, but considers nonlinear observers including the extended Kalman filter and the high-gain observer, as well as statistical estimators.

Part VI considers two important classes of chemical processes: reactors and distillation columns. In the previous parts, linear identification and linear control are applied to the chemical reactor described in detail in Chap. 19. In this first chapter of Part VI, the use of geometric nonlinear control, based on the knowledge model of the process or coupled with a state observer, is explained for a chemical reactor and a biological reactor. It must be noticed that this is the same simulated chemical reactor that was used for identification of a linear model and several discrete linear control methods. Chapter 20 sweeps the control methods used in distillation since the 1970s until our epoch that is marked by the start of industrial use of nonlinear control. Chap. 21 describes different processes and benchmarks that can be used as more or less complicated examples to test various control strategies.

Acknowledgments

Through the teaching that the author gives at the National School of Chemical Industries (ENSIC, Nancy), the contacts that he maintains with many control specialists and his knowledge of process industries, he realizes the difficulty in sharing control concepts with engineers that have a chemistry or chemical engineering background. Indeed, engineers responsible for a plant must master many disciplines, e.g. chemistry, chemical engineering, fluid mechanics and heat transfer, which allow them to understand the behaviour of their process or to explain the malfunctions that they observe. If this book is useful to students of process control who have originated from either chemical engineering or control, engineers in the plant or doctorate students beginning in process control or who want to have an overview of process control, the author's objective will have been reached.

The author wishes to thank his many colleagues of ENSIC and the Chemical Engineering Sciences Laboratory (LSGC) who have encouraged him in his work, in particular ENSIC Director A. Storck and LSGC Director D. Tondeur. He thanks his closest colleagues of the research group TASC (Treatment and Acquisition of chemical information, process Simulation and Control) of LSGC, his doctorate students, in particular in order of seniority K. Abidi, Z.L. Wang, S. Lucena, C. Gentric and M. Ben Thabet, who took interest in nonlinear control of chemical and biological processes and who shared difficulties and satisfactions, and also his other doctorate students in the hope that they

did not suffer too much during the long writing. He does not forget the staff for their pleasant collaboration climate and the Computing Department staff who helped him in particular in Linux installation. All world-developers of Latex and Linux are gratefully acknowledged.

The author also wishes to thank his colleagues who read and criticized some parts of the book: J. Ragot, Professor at the National Polytechnic Institute of Lorraine and P. Sibille, Associate Professor at Nancy I University, both members of CRAN, G. Thomas, Professor at Ecole Centrale of Lyon, S. Rohani, Professor at University of Western Ontario, U. Volk, Engineer at Shell Godorf. Let M. Alamir, Researcher at the Grenoble Control Laboratory, who accepted the heavy task of reading the whole book, find here the expression of the author's deep gratitude.

Lastly, the author dedicates this book to his children Alain and Gilles and his close relatives, who showed so much patience and understanding in the face of his being monopolized during the realization of this book.

Contents

I Continuous-Time Control 1

1 Dynamic Modelling of Chemical Processes 3
- 1.1 References . . . 3
- 1.2 Applications of Process Control . . . 3
- 1.3 Process Description from the Control Engineer's Viewpoint . . . 4
- 1.4 Model Classification . . . 7
- 1.5 State-Space Models . . . 8
- 1.6 Examples of Models in Chemical Engineering . . . 9
  - 1.6.1 Lumped-Parameter Systems . . . 10
  - 1.6.2 Distributed-Parameter Systems . . . 18
  - 1.6.3 Degrees of Freedom . . . 20
- 1.7 Process Stability . . . 20
- 1.8 Order of a System . . . 20
- 1.9 Laplace Transform . . . 21
  - 1.9.1 Linearization and Deviation Variables . . . 22
  - 1.9.2 Some Important Properties of Laplace Transformation . . . 24
  - 1.9.3 Transfer Function . . . 27
  - 1.9.4 Poles and Zeros of a Transfer Function . . . 35
  - 1.9.5 Qualitative Analysis of a System Response . . . 36
- 1.10 Linear Systems in State Space . . . 39
  - 1.10.1 General Case . . . 39
  - 1.10.2 Analog Representation . . . 40
- 1.11 Dynamic Behaviour of Simple Processes . . . 42
  - 1.11.1 First-Order Systems . . . 44
  - 1.11.2 Integrating Systems . . . 45
  - 1.11.3 Second-Order Systems . . . 46
  - 1.11.4 Higher-Order Systems . . . 50
  - 1.11.5 Process Identification in the Continuous Domain . . . 57

2 Linear Feedback Control 65
- 2.1 Design of a Feedback Loop . . . 65
  - 2.1.1 Block Diagram of the Feedback Loop . . . 65
  - 2.1.2 General Types of Controllers . . . 67
  - 2.1.3 Sensors . . . 69

2.1.4 Transmission Lines 71
2.1.5 Actuators 71
2.2 Block Diagrams, Signal-Flow Graphs, Calculation Rules 73
2.3 Dynamics of Feedback-Controlled Processes 82
2.3.1 Study of Different Actions 85
2.3.2 Influence of Proportional Action 86
2.3.3 Influence of Integral Action 90
2.3.4 Influence of Derivative Action 93
2.3.5 Summary of Controllers Characteristics 97

3 Stability Analysis **103**
3.1 Case of a System Defined by its Transfer Function 103
3.2 State-Space Analysis 104
3.2.1 General Analysis for a Continuous Nonlinear System . . 104
3.2.2 Case of a Linear Continuous System 109
3.2.3 Case of a Nonlinear Continuous System: The Polymerization Reactor 111
3.2.4 State-Space Analysis of a Linear System 117
3.3 Stability Analysis of Feedback Systems 118
3.3.1 Routh-Hurwitz Criterion 119
3.3.2 Root Locus Analysis 121
3.3.3 Frequency Method 126

4 Design of Feedback Controllers **129**
4.1 Performance Criteria 129
4.2 Transient Response Characteristics 131
4.3 Performance Criteria for Design 131
4.4 Choice of PID Controller 134
4.4.1 General Remarks 134
4.4.2 Recommendations 135
4.5 PID Controller Tuning 136
4.5.1 Tuning by Trial and Error 136
4.5.2 Sustained Oscillation Method 137
4.5.3 Relay Oscillation Method 138
4.5.4 Process Reaction Curve Method 143
4.5.5 Tuning rule of Tavakoli and Fleming for PI controllers . 146
4.5.6 Robust tuning rule for PID controllers 146
4.6 PID Improvement 147
4.6.1 PID Controller with Derivative Action on the Measured Output 147
4.6.2 Use of a Reference Trajectory 148
4.6.3 Discretized PID Controller 148
4.6.4 Anti-Windup Controller 150
4.6.5 PID Control by On−Off Action 151
4.6.6 pH Control 154
4.7 Direct Synthesis Method 159

4.8 Internal Model Control 160
4.9 Pole-Placement 166
4.9.1 Robustness of Pole-Placement Control 173
4.9.2 Unitary Feedback Controller 174
4.10 Linear Quadratic Control 175
4.10.1 Regulation Behaviour 176
4.10.2 Tracking Behaviour 176

5 Frequency Analysis **183**
5.1 Response of a Linear System to a Sinusoidal Input 183
5.1.1 Case of a First-Order Process 183
5.1.2 Note on Complex Numbers 185
5.1.3 Case of any Linear Process 186
5.1.4 Case of Linear Systems in Series 187
5.2 Graphical Representation 188
5.2.1 Bode Plot 188
5.2.2 n-th-Order System 190
5.2.3 Nyquist Plot 192
5.2.4 n-th-Order System 194
5.2.5 Black Plot 195
5.3 Characterization of a System by Frequency Analysis 195
5.4 Frequency Response of Feedback Controllers 196
5.4.1 Proportional Controller 196
5.4.2 Proportional-Integral Controller 196
5.4.3 Ideal Proportional-Derivative Controller 197
5.4.4 Proportional-Integral-Derivative Controller 198
5.5 Bode Stability Criterion 201
5.6 Gain and Phase Margins 206
5.6.1 Gain Margin 207
5.6.2 Phase Margin 208
5.7 Nyquist Stability Criterion 213
5.8 Closed-Loop Frequency Response 215
5.9 Internal Model Principle 220
5.10 Robustness 221
5.11 Summary for Controller Design 237

6 Improvement of Control Systems **239**
6.1 Compensation of Time Delay 239
6.2 Inverse Response Compensation 241
6.3 Cascade Control 244
6.4 Selective Control 249
6.5 Split-Range Control 250
6.6 Feedforward Control 251
6.6.1 Generalities 251
6.6.2 Application in Distillation 252
6.6.3 Synthesis of a Feedforward Controller 253

6.6.4 Realization of a Feedforward Controller 255
6.6.5 Feedforward and Feedback Control 256
6.7 Ratio Control 258

7 State Representation, Controllability and Observability 261
7.1 State Representation 261
7.1.1 Monovariable System 261
7.1.2 Multivariable System 263
7.2 Controllability 264
7.3 Observability 268
7.4 Realizations 271
7.5 Remark on Controllability and Observability in Discrete Time . 275

II Multivariable Control 277

8 Multivariable Control by Transfer Function Matrix 279
8.1 Introduction 279
8.2 Representation of a Multivariable Process by Transfer Function Matrix 279
8.3 Stability Study 282
8.3.1 Smith-McMillan Form 282
8.3.2 Poles and Zeros of a Transfer Function Matrix 282
8.3.3 Generalized Nyquist Criterion 282
8.3.4 Characteristic Loci 283
8.3.5 Gershgorin Circles 284
8.3.6 Niederlinski Index 286
8.4 Interaction and Decoupling 286
8.4.1 Decoupling for a 2×2 System 286
8.4.2 Disturbance Rejection 291
8.4.3 Singular Value Decomposition 291
8.4.4 Relative Gain Array 292
8.4.5 Gershgorin Circles and Interaction 298
8.5 Multivariable Robustness 298
8.6 Robustness Study of a 2×2 Distillation Column 303
8.6.1 Simplified Decoupling Analysis 303
8.6.2 Ideal Decoupling Analysis 304
8.6.3 One-Way Decoupling Analysis 304
8.6.4 Comparison of the Three Previous Decouplings 305
8.7 Synthesis of a Multivariable Controller 305
8.7.1 Controller Tuning by the Largest Modulus Method . . . 306
8.7.2 Controller Tuning by the Characteristic Loci Method . 306
8.8 Discrete Multivariable Internal Model Control 307

III Discrete-Time Identification 311

9 Discrete-Time Generalities and Basic Signal Processing 313
9.1 Fourier Transformation and Signal Processing 314
9.1.1 Continuous Fourier Transform 314
9.1.2 Discrete Fourier Transform 320
9.1.3 Stochastic Signals 322
9.1.4 Stochastic Stationary Signals 324
9.1.5 Summary 325
9.2 Sampling 329
9.2.1 D/A and A/D Conversions 329
9.2.2 Choice of Sampling Period 330
9.3 Filtering 336
9.3.1 First-Order Filter 337
9.3.2 Second-Order Filter 338
9.3.3 Moving Average Filter 339
9.3.4 Fast Transient Filter 339
9.4 Discrete-Time and Finite-Differences Models 340
9.5 Different Discrete Representations of a System 342
9.5.1 Discrete Representation: z-transform 342
9.5.2 Conversion of a Continuous Description in Discrete Time 362
9.5.3 Operators 365

10 Identification Principles 371
10.1 System Description 371
10.1.1 System without Disturbance 371
10.1.2 Disturbance Representation 372
10.2 Non-Parametric Identification 373
10.2.1 Frequency Identification 373
10.2.2 Identification by Correlation Analysis 374
10.2.3 Spectral Identification 376
10.3 Parametric Identification 379
10.3.1 Prediction Principles 379
10.3.2 One-Step Prediction 380
10.3.3 p-Step Predictions 385

11 Models and Methods for Parametric Identification 387
11.1 Model Structure for Parametric Identification 387
11.1.1 Linear Models of Transfer Functions 387
11.1.2 Models for Estimation in State Space 397
11.2 Models of Time-Varying Linear Systems 406
11.3 Linearization of Nonlinear Time-Varying Models 407
11.4 Principles of Parametric Estimation 407
11.4.1 Minimization of Prediction Errors 407
11.4.2 Linear Regression and Least Squares 409
11.4.3 Maximum Likelihood Method 412

11.4.4 Correlation of Prediction Errors with Past Data 415
11.4.5 Instrumental Variable Method 416

12 Parametric Estimation Algorithms **421**
12.1 Linear Regression and Least Squares 421
12.2 Gradient Methods 423
12.2.1 Gradient Method Based on a Priori Error 424
12.2.2 Gradient Method Based on a Posteriori Error 427
12.3 Recursive Algorithms 429
12.3.1 Simple Recursive Least Squares 429
12.3.2 Recursive Extended Least Squares 438
12.3.3 Recursive Generalized Least Squares 438
12.3.4 Recursive Maximum Likelihood 439
12.3.5 Recursive Prediction Error Method 440
12.3.6 Instrumental Variable Method 443
12.3.7 Output Error Method 444
12.4 Algorithm Robustification 444
12.5 Validation 447
12.6 Input Sequences for Identification 448
12.7 Identification Examples 450
12.7.1 Academic Example of a Second-Order System 450
12.7.2 Identification of a Simulated Chemical Reactor 456

IV Discrete Time Control **461**

13 Digital Control **463**
13.1 Pole-Placement Control 463
13.1.1 Influence of Pole Position 463
13.1.2 Control Synthesis by Pole-Placement 463
13.1.3 Relation between Pole-Placement and State Feedback . 470
13.1.4 General Pole-Placement Design 474
13.1.5 Digital PID Controller 481
13.2 Discrete Internal Model Control 483
13.3 Generalities in Adaptive Control 488

14 Optimal Control **493**
14.1 Introduction 493
14.2 Problem Statement 494
14.3 Variational Method in the Mathematical Framework 496
14.3.1 Variation of the Criterion 497
14.3.2 Variational Problem without Constraints, Fixed Boundaries 499
14.3.3 Variational Problem with Constraints, General Case . . 499
14.3.4 Hamilton-Jacobi Equation 502
14.4 Optimal Control 504
14.4.1 Variational Methods 504

14.4.2 Variation of the Criterion 505
14.4.3 Euler Conditions 506
14.4.4 Weierstrass Condition and Hamiltonian Maximization 508
14.4.5 Hamilton-Jacobi Conditions and Equation 509
14.4.6 Maximum Principle 511
14.4.7 Singular Arcs 514
14.4.8 Numerical Issues 521
14.5 Dynamic Programming 525
14.5.1 Classical Dynamic Programming 525
14.5.2 Hamilton-Jacobi-Bellman Equation 529
14.6 Linear Quadratic Control 530
14.6.1 Continuous-Time Linear Quadratic Control 531
14.6.2 Linear Quadratic Gaussian Control 541
14.6.3 Discrete-Time Linear Quadratic Control 547

15 Generalized Predictive Control 555
15.1 Interest in Generalized Predictive Control 555
15.2 Brief Overview of Predictive Control Evolution 556
15.3 Simple Generalized Predictive Control 557
15.3.1 Theoretical Presentation 557
15.3.2 Numerical Example : Generalized Predictive Control of a Chemical Reactor 561
15.3.3 GPC seen as a Pole-Placement 562
15.4 Generalized Predictive Control with Multiple Reference Model 564
15.4.1 Theoretical Presentation 564
15.4.2 Numerical Example: Generalized Predictive Control with Performance Model of a Chemical Reactor 566
15.5 Partial State Reference Model Control 568
15.6 Generalized Predictive Control of a Chemical Reactor 568

16 Model Predictive Control 575
16.1 A General View of Model Predictive Control 575
16.2 Linear Model Predictive Control 580
16.2.1 In the Absence of Constraints 580
16.2.2 In the Presence of Constraints 581
16.2.3 Short Description of IDCOM 581
16.2.4 Dynamic Matrix Control (DMC) 581
16.2.5 Quadratic Dynamic Matrix Control (QDMC) 589
16.2.6 State-Space Formulation of Dynamic Matrix Control . . 590
16.2.7 State-Space Linear Model Predictive Control as OBMPC 591
16.2.8 State-Space Linear Model Predictive Control as General Optimization 594
16.3 Nonlinear Model Predictive Control 595

16.3.1 Nonlinear Quadratic Dynamic Matrix Control (NLQDMC) 595
16.3.2 Other Approaches of Nonlinear Model Predictive Control 597
16.4 Model Predictive Control of a FCC 600
16.4.1 FCC Modelling 600

V Nonlinear Control 617

17 Nonlinear Geometric Control 619
17.1 Some Linear Notions Useful in Nonlinear Control 620
17.1.1 Influence of a Coordinate Change in Linear Control . . 620
17.1.2 Relative Degree 621
17.1.3 Normal Form and Relative Degree 622
17.1.4 Zero Dynamics 624
17.1.5 Static State Feedback 625
17.1.6 Pole-Placement by Static State Feedback 625
17.1.7 Input-Output Pole-Placement 627
17.2 Monovariable Nonlinear Control 627
17.2.1 Some Notions of Differential Geometry 627
17.2.2 Relative Degree of a Monovariable Nonlinear System . . 629
17.2.3 Frobenius Theorem 630
17.2.4 Coordinates Change 632
17.2.5 Normal Form 633
17.2.6 Controllability and Observability 634
17.2.7 Principle of Feedback Linearization 635
17.2.8 Exact Input-State Linearization for a System of Relative Degree Equal to n 635
17.2.9 Input-Output Linearization of a System with Relative Degree r Lower than or Equal to n 638
17.2.10 Zero Dynamics 639
17.2.11 Asymptotic Stability 642
17.2.12 Tracking of a Reference Trajectory 644
17.2.13 Decoupling with Respect to a Disturbance 644
17.2.14 Case of Nonminimum-Phase Systems 646
17.2.15 Globally Linearizing Control 646
17.3 Multivariable Nonlinear Control 647
17.3.1 Relative Degree 648
17.3.2 Coordinate Change 649
17.3.3 Normal Form 649
17.3.4 Zero Dynamics 650
17.3.5 Exact Linearization by State Feedback and Diffeomorphism 651
17.3.6 Nonlinear Control Perfectly Decoupled by Static State Feedback 651
17.3.7 Obtaining a Relative Degree by Dynamic Extension . . 653
17.3.8 Nonlinear Adaptive Control 654

17.4 Applications of Nonlinear Geometric Control 654

18 State Observers **657**
18.1 Introduction 657
18.1.1 Indirect Sensors 658
18.1.2 Observer Principle 658
18.2 Parameter Estimation 659
18.3 Statistical Estimation 659
18.3.1 About the Data 660
18.3.2 Principal Component Analysis 660
18.3.3 Partial Least Squares 662
18.4 Observers 663
18.4.1 Luenberger Observer 663
18.4.2 Linear Kalman Filter 664
18.4.3 Extended Kalman Filter (EKF) in Continuous-Discrete Form 664
18.4.4 Globally Linearizing Observer 666
18.4.5 High-Gain Observer 667
18.4.6 Moving state estimation 669
18.5 Conclusion 673

VI Applications to Processes **679**

19 Nonlinear Control of Reactors with State Estimation **681**
19.1 Introduction 681
19.2 Chemical Reactor 681
19.2.1 Model of the Chemical Reactor 682
19.2.2 Control Problem Setting 683
19.2.3 Control Law 684
19.2.4 State Estimation 687
19.2.5 Simulation Results 688
19.3 Biological Reactor 692
19.3.1 Introduction 692
19.3.2 Dynamic Model of the Biological Reactor 693
19.3.3 Synthesis of the Nonlinear Control Law 695
19.3.4 Simulation Conditions 697
19.3.5 Simulation Results 699
19.3.6 Conclusion 699

20 Distillation Column Control **705**
20.1 Generalities for Distillation Columns Behaviour 705
20.2 Dynamic Model of the Distillation Column 708
20.3 Generalities for Distillation Column Control 712
20.4 Different Types of Distillation Column Control 714
20.4.1 Single-Input Single-Output Control 714
20.4.2 Dual Decoupling Control 714

20.4.3 The Column as a $\mathbf{5} \times \mathbf{5}$ System 716
20.4.4 Linear Digital Control 718
20.4.5 Model Predictive Control 720
20.4.6 Bilinear Models 721
20.4.7 Nonlinear Control 724
20.5 Conclusion 728

21 Examples and Benchmarks of Typical Processes 733
21.1 Single-Input Single-Output Processes 733
21.1.1 Description by Transfer Functions 733
21.1.2 Description by a State-Space Knowledge Model 734
21.1.3 Description by a Linear State-Space Model 740
21.2 Multivariable Processes 740
21.2.1 Matrices of Continuous Transfer Functions 740
21.2.2 Description by a Linear State-Space Model 743
21.2.3 Description by a State-Space Knowledge Model 745
21.2.4 Continuous State-Space Models 747

Part I

Continuous-Time Control

Chapter 1

Dynamic Modelling of Chemical Processes

1.1 References

Many textbooks and research books are available. When a precise reference is given for a textbook which describes a given topic particularly well, the present author has tried not to forget his colleagues and mentions them at the end of the concerned chapter. When general textbooks can be recommended for different points, they are mentioned at the end of this first chapter to avoid too much repetition. It is impossible to cite all textbooks and the fact that some are not cited does not mean that they are of lower value. Of course, research papers are referenced in the concerned chapter.

1.2 Applications of Process Control

A chemical plant represents a complex arrangement of different units (reactors; separation units such as distillation, absorption, extraction, chromatography and filtration; heat exchangers; pumps; compressors; tanks; ...). These units must be either maintained close to their steady states for continuous operation or follow optimal trajectories for batch operation.

The engineers in charge of a plant must ensure quantitative and qualitative product specifications and economic performance while meeting health, safety and environmental regulations.

The task of a control system is to ensure the stability of the process, to minimize the influence of disturbances and perturbations and to optimize the overall performance. These objectives are achieved by maintaining some variables (temperature, pressure, concentration, position, speed, quality, ...) close to their desired values or using set points which can be fixed or time-dependent.

When a chemical engineer designs a process control system, he or she must first study the process and determine its characteristics. The process variables

are classified as inputs, outputs, and states.

Subsequently, a process model with varying degrees of complexity (according to the ultimate use of the model) is derived. For an existing process, a black box model (where coefficients have no physical meaning) may be developed by system identification techniques, often with little effort. On the other hand, a physical model based on first principles involves a great deal of engineering effort.

The developed model is in general verified in an off-line manner and independently of the control scheme. Then, the model is used together with the chosen control scheme to check the process response to set point and disturbances variations.

The control scheme can be a simple proportional controller as well as a much more sophisticated algorithm, e.g. if nonlinear control based on a physical model of the process is utilized.

The outputs of the simulation model used for control can be compared to those of the real process or to those from an accurate model. System identification (Fig. 1.1) allows the engineer to estimate the parameters of the model or to evaluate the performances of the control law.

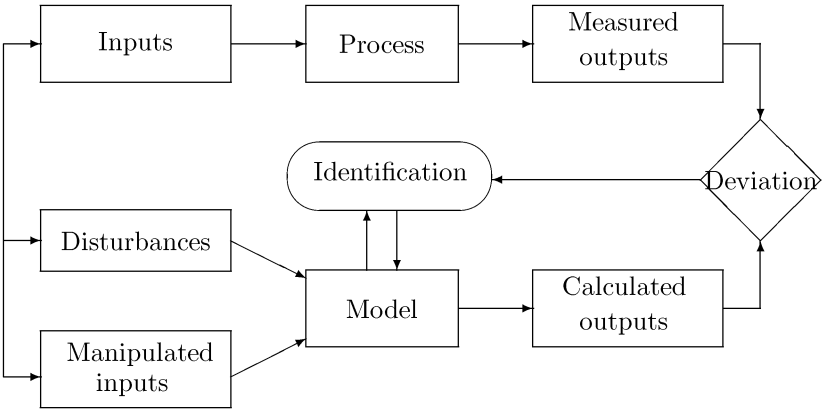


Figure 1.1: Design of a process model for control

1.3 Process Description from the Control Engineer's Viewpoint

The control engineer considers the process to be controlled as a dynamic system having inputs, outputs and internal variables called the state variables. His/her classification is different from the process engineer's point of view. Indeed, for a process engineer, inputs are essentially physical streams (such as a feed pipe) delivering material and energy to the process, possibly information such as electrical signals, and outputs are similarly physical streams, withdrawing

materials (the effluent stream to the downstream processing unit) and energy from the process.

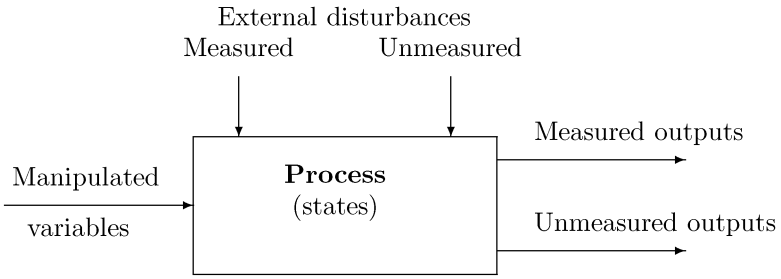


Figure 1.2: Input-output block diagram representation of a process

From the control engineer's point of view, variables associated with a process (flow rate, concentration, temperature, pressure, quality, ...) are generally considered as signals transferring information. These variables are divided into two groups (Fig. 1.2):

- Inputs which represent the influence of environment on the process: these variables affect the process and thus modify its behaviour.
- Outputs which represent the process influence on environment: these variables represent the link with the outside. They should be maintained close to their set points.

Input and output variables are linked by state variables (see state representation) which are internal to the process and help to describe the evolution of the process with time. Any modification of the inputs affects dynamically the process states, which in turn influence algebraically the process outputs.
Input variables are divided into:

- Control variables or manipulated variables which can be adjusted freely by the operator or by regulatory means. For example, the position of a valve stem determines the flow rate through the valve. In this case, the input or manipulated variable will be the valve stem position or the flow rate in the pipe which is directly related to the valve position.
- Disturbances include all other inputs which are not set by the operator or a controller. For example, in the case of temperature control in a building, outside climate temperature and humidity changes are considered as disturbances which affect the controlled variable that is the inner temperature.

Output variables are divided into:

- Measured variables, whose values are known using direct on-line measurement by a sensor.
- Unmeasured variables, which may be estimated or inferred using an indirect or secondary measurement. This estimation of unmeasured variables by means of a model and other measurements constitutes a soft sensor (Chap. 18).

Figure 1.2 represents an input-output block diagram of a process.

A system is called single variable or single-input single-output (SISO) if it has only one input and one output. It is called multivariable or multi-input multi-output (MIMO) if it has several inputs and several outputs. In general, the number of inputs is larger than the number of outputs.

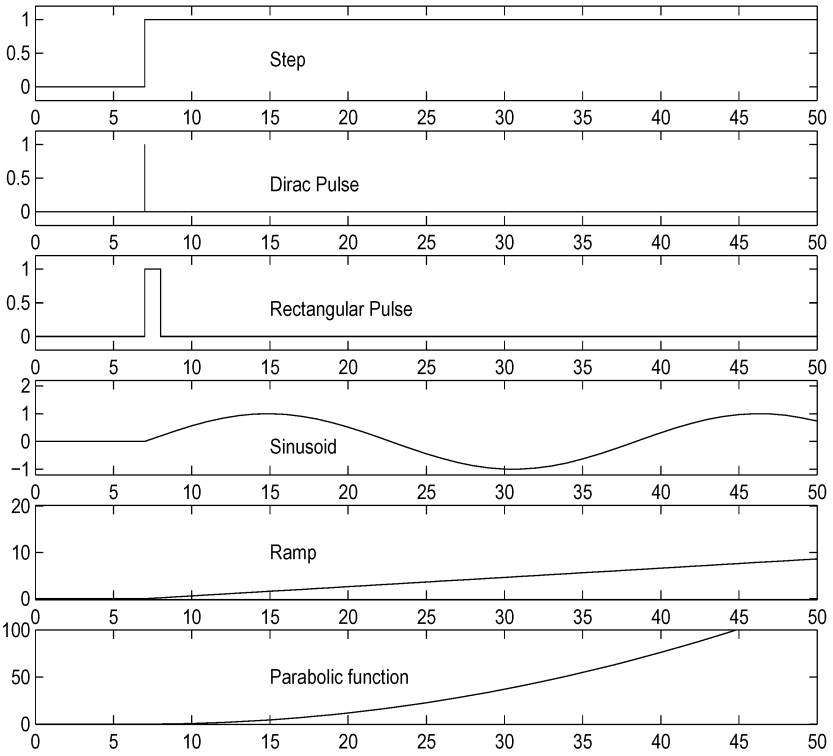


Figure 1.3: Different types of inputs starting at time $t = 7$ after steady state

In order to study the behaviour of a process, the process inputs are generally varied by some simple and standard functions and the response of the process is monitored in each case. Apart from the disturbances, which can take any form, the standard input functions f (Fig. 1.3) are:

- A step function: a unit step function is defined as $f = 1$ if $t > 0$, $f = 0$ if $t \leq 0$. Its response is called a step response.
- An impulse function: a unit impulse function is defined as: $f = \delta$ (theoretical Dirac). Its response is called an impulse response.
- A sinusoidal function: $f = a\cos(\omega t + \phi)$. Its response is referred to as a frequency response.
- A ramp: $f = kt$. This determines the behaviour of the process output to an input with constant rate of change (constant velocity).
- A parabolic function: $f = kt^2$. This is used whenever the response to a

constant acceleration is desired.

The inputs to a controlled physical system can be further classified as disturbances (loads) (for regulatory control) or set point variations (for set point tracking or servo-control).

1.4 Model Classification

Models can be classified with respect to different user-specified criteria. In the steady-state models, the time derivatives of the state variables are set to zero ($d/dt = 0$). In the dynamic models which describe the transient behaviour of the process, the process variables are time-dependent ($f(t, x, u)$). In process control applications, the models must be dynamic in order to represent the process variations with respect to time.

Dynamic models can be of two kinds: deterministic, in which it is assumed that all the variables are perfectly known at a given instant of time, or probabilistic (stochastic) models, which make use of probability distributions to account for the variations and uncertainties associated with the process and its variables.

Dynamic models can be continuous when the function $f(t, x, u)$ describing the process is continuous or discrete with respect to time (variables are only known at regular time intervals).

A model can be developed using merely the process input-output data series without physical knowledge of the process. This type of model is referred to as a black-box or behavioural model (such as neural networks). At the other extreme, a model may be developed from the application of first principles (conservation laws) to the process. Such models are called phenomenological or knowledge-based.

The knowledge-based models are further classified as:

- Lumped-parameter models in which the state variables have no spatial dependence, and therefore the models consist of ordinary differential equations (e.g. the classical continuous stirred tank reactor).
- Distributed-parameter models, in which the state variables are position-dependent and the models take the form of partial differential equations (e.g. a tubular reactor). Very often, a distributed-parameter model is discretized (division of a tubular reactor into n continuous stirred tank reactors) so as to transform the partial differential equations into a set of ordinary differential equations that are more tractable mathematically.

Furthermore, models can be linear or nonlinear. A process model is linear if all variables describing the process appear linearly in the equations. It is nonlinear in the opposite case. The advantage of linear models is that they can be easily transformed by mathematical mappings and their mathematical behaviour is well known.

Among the many forms of models that will be used in this book, the transfer functions (in the continuous s Laplace or discrete z domains) and the state-space models are of special significance.

A key point which must be remembered is that a model, however sophisticated it may be, remains an approximation of the real process. It will often differ according to the pursued objective.

1.5 State-Space Models

Generally, a state-space multivariable system is modelled by a set of algebraic and differential equations of the form

$$\begin{cases} \dot{\boldsymbol{x}} = & \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) \\ \boldsymbol{y} = & \boldsymbol{h}(\boldsymbol{x}, t) \end{cases} \tag{1.1}$$

where $\boldsymbol{x}$ is the state vector of dimension n, $\boldsymbol{u}$ is the input vector (or control variables vector) of dimension n_u and $\boldsymbol{y}$ is the output vector of dimension n_y (in general, $n_u \geq n_y$). The state $\boldsymbol{x}$ of the system at any time t is described by a set of n differential equations, which are often non-linear. The states $\boldsymbol{x}(t)$ depend only on initial conditions at t_0 and on inputs $\boldsymbol{u}(t)$ between t_0 and t. At the initial time, the state variables are the initial conditions and later they represent the evolution of the state of the system.

A remarkable characteristic of the non-linear dynamic knowledge-based models in chemical engineering is that they are described by differential equations, either ordinary or partial, which are usually affine with respect to the input vector. Thus, the most common form of a non-linear state-space model is

$$\begin{cases} \dot{\boldsymbol{x}} & = \boldsymbol{f}(\boldsymbol{x}, t) + \boldsymbol{g}(\boldsymbol{x}, t)\, \boldsymbol{u} \\ \boldsymbol{y} & = \boldsymbol{h}(\boldsymbol{x}, t) \end{cases} \tag{1.2}$$

An important class is linear state-space models of the form

$$\begin{cases} \dot{\boldsymbol{x}}(t) & = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \\ \boldsymbol{y}(t) & = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) \end{cases} \tag{1.3}$$

which includes any process model described by a set of n linear ordinary differential equations. $\boldsymbol{A}$ is the state matrix of dimension $(n \times n)$, $\boldsymbol{B}$ is the control matrix of dimension $(n \times n_u)$, $\boldsymbol{C}$ is the output matrix of dimension $(n_y \times n)$, and $\boldsymbol{D}$ is the coupling matrix of dimension $(n_y \times n_u)$ which is very often equal to zero. When $\boldsymbol{D}$ is different from zero, it is said that the output is directly driven by the input.

In some cases, it is necessary to model a linear system under the following generalized state-space representation or descriptor form

$$\begin{cases} \boldsymbol{E}\dot{\boldsymbol{x}}(t) & = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \\ \boldsymbol{y}(t) & = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) \end{cases} \tag{1.4}$$

where $\boldsymbol{E}$ is called the descriptor matrix of dimension $(n \times n)$. If $\boldsymbol{E}$ is invertible, the system can be rewritten under the form (1.3). If $\boldsymbol{E}$ is singular, this corresponds to an differential-algebraic system.

The linearization of a set of differential equations of the most general form

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) \tag{1.5}$$

around a steady-state operating level $\boldsymbol{x}_0$ for a nominal input $\boldsymbol{u}_0$ can be achieved by the following Taylor series expansion

$$\dot{x}_i = f_i(\boldsymbol{x}_0, \boldsymbol{u}_0, t) + \sum_{j=1}^{n} \left(\frac{\partial f_i}{\partial x_j} \right)_{\boldsymbol{x}=\boldsymbol{x}_0, \boldsymbol{u}=\boldsymbol{u}_0} \delta x_j + \sum_{k=1}^{n_u} \left(\frac{\partial f_i}{\partial u_k} \right)_{\boldsymbol{x}=\boldsymbol{x}_0, \boldsymbol{u}=\boldsymbol{u}_0} \delta u_k \tag{1.6}$$

with: $\delta x_j = x_j - x_{j,0}$, $\delta u_k = u_k - u_{k,0}$ where $x_{j,0}$ is the j component of steady-state vector $\boldsymbol{x}_0$ and $u_{k,0}$ is the k component of the nominal input vector $\boldsymbol{u}_0$. δ indicates any deviation with respect to the nominal operating condition, thus δx and δu are respectively deviations of the state and of the input with respect to the steady state of the process. One gets $\dot{x}_i = \delta \dot{x}_i + \dot{x}_{i,0}$ and $\dot{x}_{i,0} = f_{i,0}$, so that the following set of linear ordinary differential equations is obtained

$$\delta \dot{x}_i = \sum_{j=1}^{n} \left(\frac{\partial f_i}{\partial x_j} \right)_{\boldsymbol{x}=\boldsymbol{x}_0, \boldsymbol{u}=\boldsymbol{u}_0} \delta x_j + \sum_{k=1}^{n_u} \left(\frac{\partial f_i}{\partial u_k} \right)_{\boldsymbol{x}=\boldsymbol{x}_0, \boldsymbol{u}=\boldsymbol{u}_0} \delta u_k \tag{1.7}$$

which can be easily written in a more condensed matrix form similar to Eq. (1.3). Similarly

$$\delta y_i = \sum_{j=1}^{n} \left(\frac{\partial h_i}{\partial x_j} \right)_{\boldsymbol{x}=\boldsymbol{x}_0, \boldsymbol{u}=\boldsymbol{u}_0} \delta x_j + \sum_{k=1}^{n_u} \left(\frac{\partial h_i}{\partial u_k} \right)_{\boldsymbol{x}=\boldsymbol{x}_0, \boldsymbol{u}=\boldsymbol{u}_0} \delta u_k \tag{1.8}$$

The form of Eqs. (1.7) and (1.8) is referred to as the linearized state-space model with the matrices of the linear state-space model indicated by their current element

$$\boldsymbol{A} = \left[\frac{\partial f_i}{\partial x_j} \right] \quad ; \quad \boldsymbol{B} = \left[\frac{\partial f_i}{\partial u_j} \right] \quad ; \quad \boldsymbol{C} = \left[\frac{\partial h_i}{\partial x_j} \right] \quad ; \quad \boldsymbol{D} = \left[\frac{\partial h_i}{\partial u_j} \right] \tag{1.9}$$

$\boldsymbol{A}$ and $\boldsymbol{B}$ are respectively the Jacobian matrices of $\boldsymbol{f}$ with respect to $\boldsymbol{x}$ and $\boldsymbol{u}$, while $\boldsymbol{C}$ and $\boldsymbol{D}$ are respectively the Jacobian matrices of $\boldsymbol{h}$ with respect to $\boldsymbol{x}$ and $\boldsymbol{u}$.

Notions of controllability and observability in the state space will be specially studied in Chap. 7.

1.6 Examples of Models in Chemical Engineering

With the help of certain classical examples in chemical engineering, we will show how various chemical processes can be described in the state-space form.

1.6.1 Lumped-Parameter Systems

In lumped-parameter systems, the process variables depend only on time, which is the independent variable. The dynamic behaviour of the process is then described by a set of ordinary differential equations.

A Surge Tank

Consider a cylindrical tank fed by an incompressible liquid (Fig. 1.4) at a varying flow rate F_0 (m^3/s). The exit flow rate F is also time-dependent.

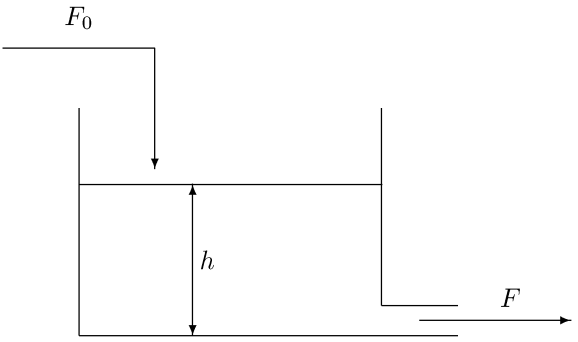


Figure 1.4: A surge tank with varying level

At steady-state, the level in the tank is constant and the mass balance requires equality of inlet and outlet mass flow rates

$$F_0\rho_0 = F\rho \tag{1.10}$$

where ρ is the fluid density.
In transient regime, the liquid level h in the tank varies with time.

The general mass balance formulated as

(inlet mass / unit time) = (outlet mass / unit time) + (time rate of change of mass in the system)

gives the dynamic model of the tank

$$F_0\,\rho_0 = F\,\rho + \frac{d(\rho V)}{dt} \tag{1.11}$$

This is the state-space equation of the process.

If the liquid density is constant ($\rho = \rho_0$) and the tank cross-sectional area S does not depend on the liquid level, the previous mass balance will become

$$\frac{dh}{dt} = F_0/S - F/S \tag{1.12}$$

One notices that only the liquid level in the tank, the controlled variable, appears in the derivative. With the section area S being a constant parameter,

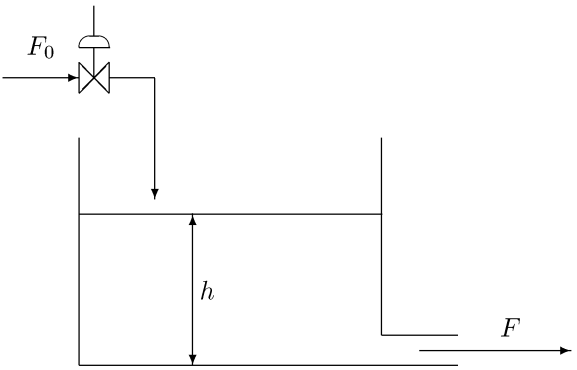


Figure 1.5: Surge tank with varying level with valve on the inlet stream

this ordinary differential equation is linear. Equation (1.12) can be considered as the fundamental model of a level control system.

Assuming that a valve is placed on the inlet pipe (Fig. 1.5), the inlet flow rate F_0 will become the control (manipulated) variable of the system. In state space, the system is single-input single-output (SISO), with an input $u = F_0$, an output $y = h$ and only one state: $x = h$. The state-space model is

$$\begin{aligned} \dot{x} &= u/S - F/S \\ y &= x \end{aligned} \tag{1.13}$$

In this linear model, the cross-sectional area S is a constant parameter, and the flow rate F is an external disturbance acting on the system.

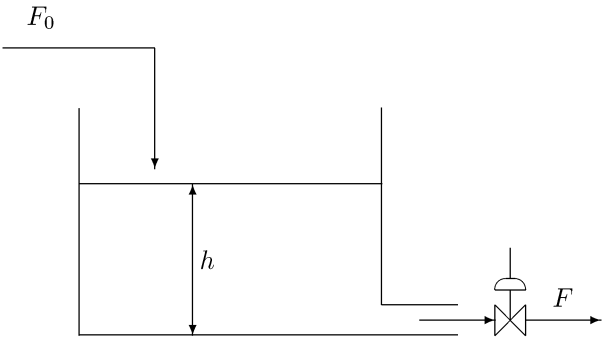


Figure 1.6: Surge tank with varying level with valve on the outlet stream

In the case where the valve is on the outlet stream (Fig. 1.6), the manipulated input is the outlet flow rate F, so that the state-space model becomes

$$\begin{aligned} \dot{x} &= F_0/S - u/S \\ y &= x \end{aligned} \tag{1.14}$$

In this case, the flow rate F_0 is a disturbance.

An Isothermal Continuous Chemical Reactor

The cylindrical reactor (Fig. 1.7) is assumed to be perfectly mixed, i.e. the temperature T, the concentration of any given species, the pressure, etc. are identical at any location of the reactor and keep their value in the effluent stream. The reactor is fed by two streams, one having a flow rate F_0 containing the reactant A, and the other one an inert stream with a flow rate F_1. Both streams have their temperature equal to that of the reactor contents. Furthermore, we will assume that the heat of reaction is negligible so that there is no need to write the energy balance.

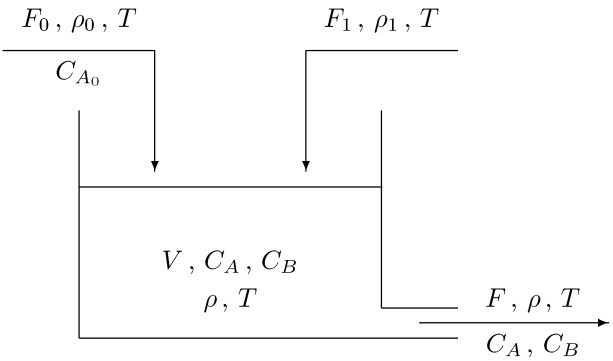


Figure 1.7: An isothermal continuous stirred tank reactor (CSTR)

The overall mass balance equation is similar to that of the previous tank

$$\frac{d(\rho\, h)}{dt} = F_0\, \rho_0/S + F_1\, \rho_1/S - F\, \rho/S \tag{1.15}$$

and we assume that the densities in different streams are equal.

The balance for component A in transient regime is given by the continuity equation for A:

(rate of A entering) = (rate of A exiting) − (rate of A produced) + (rate of accumulation of A),

giving

$$F_0\, C_{A_0} = F\, C_A - R_A\, V + \frac{d(V\, C_A)}{dt} \tag{1.16}$$

noting that $F'_{A0} = F_0\, C_{A_0}$ is the molar flow rate (mol/s) of component A in the inlet stream, and similarly F'_A is the outlet molar flow rate. Equation (1.16) can be written as

$$F'_{A0} = F'_A - R_A\, V + \frac{d(V\, C_A)}{dt} \tag{1.17}$$

The term R_A represents the number of moles of A produced per unit volume and unit time; it can be called the production rate of A (Levenspiel, 1999;

Villermaux, 1982). When R reactions designated by i occur simultaneously, the rate of production of a component A_j is equal to

$$R_j = \sum_{i=1}^{R} \nu_{ij} r_i \tag{1.18}$$

where each reaction rate r_i is in general positive. The stoichiometric coefficient $\nu_{ij} > 0$ if A_j is produced by reaction i, and $\nu_{ij} < 0$ if A_j is consumed by reaction i. Note that this definition of reaction can be applied to either a continuous reactor, a batch reactor ($F_0 = F = 0$) or a fed-batch reactor($F = 0$ and $F_0 \neq 0$). The chemical advancement ξ (dimension: mol) and the generalized yield χ (without dimension) are defined such that:

- For a closed reactor:
  Taking n_0 as the total number of moles of reacting species present at a reference state, in general the initial state, the number of moles of a component A_j at any other state is equal to

$$n_j = n_{j0} + \sum_{i=1}^{R} \nu_{ij}\, \xi_i = n_{j0} + n_0 \sum_{i=1}^{R} \nu_{ij}\, \chi_i \qquad \text{with: } n_0 = \sum_j n_{j0} \tag{1.19}$$

  given the rate of reaction i:

$$r_i = \frac{1}{V} \frac{d\xi_i}{dt} = \frac{n_0}{V} \frac{d\chi_i}{dt} \tag{1.20}$$

- For an open continuous stirred reactor:
  The reference is in general the inlet molar flow rate F'_0 including all entering reacting species j present in the reference state. So the molar flow rate F'_j of a component A_j at any point of the system is equal to

$$F'_j = F'_{j0} + \sum_{i=1}^{R} \nu_{ij}\, \dot{\xi}_i = F'_{j0} + F'_0 \sum_{i=1}^{R} \nu_{ij}\, \chi_i \qquad \text{with: } F'_0 = \sum_j F'_{j0} \tag{1.21}$$

  given the rate of reaction i:

$$r_i = \frac{F'_0}{V} (\chi_{i,out} - \chi_{i,in}) = \frac{F_0\, C_0}{V} (\chi_{i,out} - \chi_{i,in}) \tag{1.22}$$

  where *in* denotes the inlet stream and *out* the exit stream. C_0 is the total concentration in the reference state of all the constituents. The residence time τ is equal to

$$\tau = \frac{V}{F_0} = \frac{C_0\, (\chi_{i,out} - \chi_{i,in})}{r_i} \tag{1.23}$$

  In the general case of R simultaneous reactions, Equation (1.16) becomes

$$S \frac{d(h\, C_A)}{dt} = F_0\, C_{A_0} - F\, C_A + \sum_{i=1}^{R} \nu_{iA} r_i\, V \tag{1.24}$$

In the case where only one first-order chemical reaction occurs: $A \rightarrow B$, the reaction rate r_A is equal to $r_A = k\,C_A$, and the production rate of A is equal to $R_A = -r_A$, as $\nu_{iA} = -1$. Equation (1.16) becomes

$$F_0\,C_{A_0} = F\,C_A + k\,C_A\,V + \frac{d(V\,C_A)}{dt} \tag{1.25}$$

which can be transformed into

$$\frac{dC_A}{dt} = \frac{F_0}{S\,h}\,(C_{A_0} - C_A) - \frac{F_1}{S\,h}\,C_A - k\,C_A \tag{1.26}$$

This balance will be used as the fundamental model for the control of the concentration C_A.

The differential equation describing the variations in the concentration C_A is in general nonlinear, since the inlet and outlet flow rates F_0 and F are time-varying and the reaction rate can be a complicated function with respect to concentration. Frequently, a chemical reaction is either endothermic or exothermic. Therefore, an energy balance equation should be added to the previous differential equations.

Assuming that the inputs are the inlet volumetric flow rate F_1 and inlet concentration C_{A_0}, the control vector is $\boldsymbol{u} = [F_1, C_{A_0}]^T$. We wish to control the reactor level and concentration, thus the output vector is $\boldsymbol{y} = [h, C_A]^T$. The process is a 2×2 multivariable system. The state vector is chosen to be equal to $\boldsymbol{x} = [h, C_A]^T$. The state-space representation of the isothermal reactor with a first-order reaction is

$$\begin{cases} \dot{x}_1 = & F_0/S - F/S + u_1/S \\ \dot{x}_2 = & \dfrac{1}{S\,x_1}\,[F_0\,(u_2 - x_2) - x_2\,u_1] - k\,x_2 \\ y_1 = & x_1 \\ y_2 = & x_2 \end{cases} \tag{1.27}$$

This multi-input multi-output (MIMO) model is nonlinear even if we assume that the level h is perfectly regulated by an independent controller because of the differential equation describing the concentration variations. Disturbances are flow rates F_0 and F.

Frequently, when the main objective is concentration control, the influence of level variations is considered as secondary and the level can be regulated independently.

A Non-isothermal Continuous Chemical Reactor

Figure 1.8 represents the schematics of a non-isothermal continuous chemical reactor with a heating/cooling jacket. The heating/cooling medium may also be supplied by means of a coil immersed inside the reactor. It is used for cooling in cases of exothermic reaction or desired temperature decrease, and for heating in cases of endothermic reaction or desired temperature increase.

The rate of heat transfer transferred between the reacting mixture and the heating/cooling medium is $\dot{Q}$. $\dot{Q}$ is positive for heating the reacting mixture and negative in the opposite case. $\dot{Q}$ is given by

$$\dot{Q} = U\, S_{\text{ex}}\, (T_j - T) \tag{1.28}$$

where U is the overall heat transfer coefficient, S_{ex} is the available heat exchange area, T_j is the mean temperature in the jacket, and T is the temperature of the reactor.

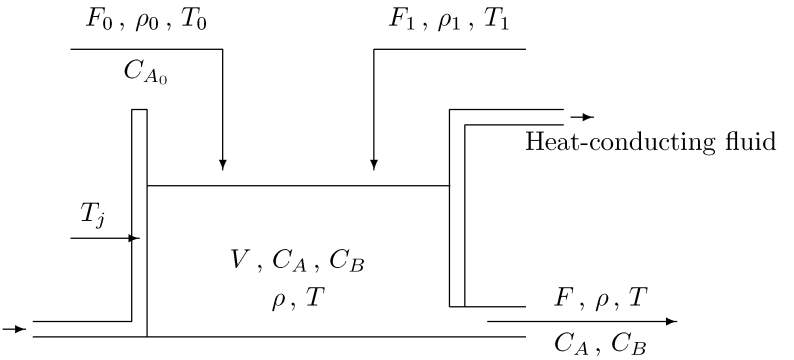


Figure 1.8: A non-isothermal continuous stirred tank chemical reactor (CSTR)

Apart from the overall mass balance and component mass balance on A, the energy balance written in general terms as

(variation of internal energy / unit time) =
(inlet enthalpy by convection / unit time) −
(outlet enthalpy by convection / unit time) +
(rate of heat transfer and mechanical energy)

must also be considered. The energy balance of the reactor is thus as follows

$$\frac{dU}{dt} = F'_{in}\, h_{in} - F'_{out}\, h_{out} + \dot{Q} \tag{1.29}$$

where F'_{in} is the total inlet molar flow rate and F'_{out} is the total outlet molar flow rate. h' is the specific molar enthalpy of each stream given by

$$h'_{in} = \sum_j x_{j,in}\, h'_{j,in} \qquad h'_{out} = \sum_j x_{j,out}\, h'_{j,out} \tag{1.30}$$

where $x_{j,in}$ and $x_{j,out}$ are the inlet and outlet mole fractions respectively and $h'_{j,in}$ and $h'_{j,out}$ are the specific enthalpies of component j in the inlet and outlet streams respectively. The specific molar enthalpy h' of a pure component can

be expressed with respect to its enthalpy of formation $\Delta H_f(T_{ref})$ at a reference temperature T_{ref} according to

$$h' = \Delta H_f(T_{ref}) + \int_{T_{ref}}^{T} C'_p \, d\tau \tag{1.31}$$

provided that there is no change of state between T_{ref} and T. If, however, there is a change of state, the corresponding latent heat of transformation should be accounted for. C'_p is the molar specific heat of the component at constant pressure.

The change of total internal energy U of the reactor contents can be expressed in terms of the specific molar internal energy by

$$\frac{dU}{dt} = \frac{d(n\,u)}{dt} = \frac{d(H - P\,V)}{dt} = \frac{d(n\,h)}{dt} - P\,\frac{dV}{dt} - V\,\frac{dP}{dt} \tag{1.32}$$

For isobaric operation and negligible pressure work, there remains

$$\frac{dU}{dt} = \frac{d(n\,h)}{dt} \tag{1.33}$$

The enthalpy change Δh_i due to reaction i is equal to

$$\Delta h_i = \sum_j \nu_{ij}\, h_j \tag{1.34}$$

which concerns only reacting and produced species, not inert ones. For R reactions, the total enthalpy contribution dh linked to the reactions is

$$dh = V dt \sum_{i=1}^{R} r_i\, \Delta h_i \tag{1.35}$$

The enthalpy of the reactor walls and its accessories such as the mixer should also be considered. Assuming that the mass of the reactor wall and its accessories is represented by m_r and the corresponding heat capacity is C_r, the overall thermal balance after some mathematical manipulation can be written as

$$(m\,C_p + m_r\,C_r)\,\frac{dT}{dt} = \sum_{in} F_{in}\,\rho_{in}\,C_{p,in}\,(T_{in} - T) + \dot{Q} - V \sum_{i=1}^{R} r_i\,\Delta h_i \tag{1.36}$$

where m is the mass of the reactor contents, C_p its mean specific heat calculated at the reactor temperature T, and the summation is carried out over all inlet streams to the reactor. Note that the reactor is perfectly mixed, therefore the exit temperature and concentration are identical to those in the reactor. The heats of reactions Δh_i are calculated at the reactor temperature.

Consider again the first-order chemical reaction $A \rightarrow B$ with heat of reaction $\Delta h_{A\rightarrow B}$, taking place in the reactor shown in Fig. 1.8. The temperature

dependency of the reaction rate is expressed by the Arrhenius equation $r = k_0 \exp(-E/(RT))\, C_A$, which is a highly nonlinear term. When applied to the reactor shown in Fig. 1.8 with its two inlet streams, Eq. (1.36) becomes (assuming that densities ρ_0, ρ_1, ρ and heat capacities are constant)

$$(m\, C_p + m_r\, C_r)\, \frac{dT}{dt} = F_0\, \rho_0\, C_{p0}\, (T_0 - T) + F_1\, \rho_1\, C_{p1}\, (T_1 - T) + \dot{Q} - V\, r\, \Delta h_{A \to B} \tag{1.37}$$

where C_{po} is the mean specific heat of the inlet stream with volumetric flow rate F_0. The densities are assumed to be identical and constant.

Notice that the differential equation describing the temperature variation is nonlinear. This model will be used for temperature control studies in the remainder of the book.

The behaviour of the chemical reactor shown in Fig. 1.8 is described by a set of three coupled ordinary differential equations.

In addition to the level and concentration in the case of the isothermal reactor, temperature must also be controlled. Moreover, it is assumed that a valve allows us to manipulate the thermal power $\dot{Q}$ introduced in the jacket. This can be performed by a heat exchanger. Therefore, the output vector is $\boldsymbol{y} = [h, C_A, T]^T$, the control vector is $\boldsymbol{u} = [F_1, C_{A_0}, \dot{Q}]^T$, and the state vector is $\boldsymbol{x} = [h, C_A, T]^T$.

In this case, the state-space model obtained from balance equations is as follows

$$\begin{cases} \dot{x}_1 = & F_0/S - F/S + u_1/S \\ \dot{x}_2 = & \dfrac{1}{S\, x_1}\, [F_0\, (u_2 - x_2) - x_2\, u_1] - k_0\, \exp(-E/(R\, x_3))\, x_2 \\ \dot{x}_3 = & \dfrac{1}{\rho\, S\, C_p\, x_1 + m_r C_r}\, [F_0\, \rho_0\, C_{p0}\, (T_0 - x_3) + u_1\, \rho_1\, C_{p1}\, (T_1 - x_3) \\ & -k_0\, \exp(-E/(R\, x_3))\, S\, x_1\, x_2 \Delta h_T + u_3] \\ y_1 = & x_1 \\ y_2 = & x_2 \\ y_3 = & x_3 \end{cases} \tag{1.38}$$

The state-space model is multi-input multi-output and nonlinear because of the two differential equations describing respectively the variations of concentration and of temperature in the reactor. Nevertheless, this model is affine with respect to manipulated variables.

In general, this model will be too complicated for control system design and analysis. If one is primarily concerned with temperature control, only the energy balance will be taken into account.

Staged Processes

A tray distillation column is composed of a finite number of stages, similar to a tray absorption column or a multi-stage mixer-settler for liquid-liquid

extraction. In a staged process, the overall mass balance, the component mass balance and the energy balance are applied to each stage separately. The full state-space model of a staged process is thus obtained by gathering the stage models and taking into account the relations between the stages and the environment. Therefore, the overall model of these processes consists of a large number of ordinary differential equations adequate for dynamic simulation but which pose difficult problems for control studies and implementation. The number of differential equations (the model order), however, can be reduced by efficient model reduction techniques to obtain approximate low-order models. The reduced model must keep the main dynamic characteristics of the original high-order model. An example of model reduction for a distillation column is given in Chap. 20.

1.6.2 Distributed-Parameter Systems

When process variables depend simultaneously on time and spatial variables, the process is described by partial differential equations.

A chemical tubular reactor, chromatography column, packed absorption or distillation column are examples of distributed-parameter systems. Only one example will be presented here: an isothermal tubular reactor (Fig. 1.9). If noticeable heat exchange occurs in the reactor, the energy balance equations must be considered as previously with heat exchange occurring at the wall.

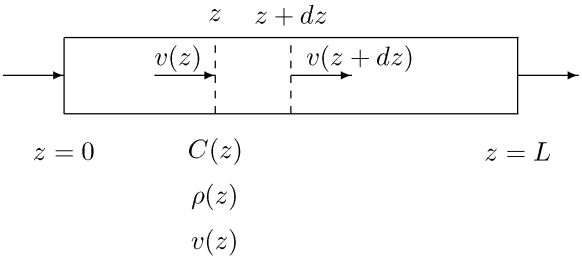


Figure 1.9: Schematics of a tubular reactor

The flow is assumed to be fully turbulent, which results in a flat velocity profile and justifies the plug flow assumption. For a plug flow reactor, the reaction rate and all the process variables are constant over a given cross-section, i.e. the radial variations are discarded and only the axial variations are considered. The chemical reaction is identical to the previous cases: $A \rightarrow B$, first-order, and the reaction rate r_A is given by $r_A = k_0 \exp(-E/(RT))\, C_A$.

The conservation principles, in the case of distributed-parameter systems, are applied to an infinitesimal volume (control volume) in which the fluid properties may be assumed constant. The shape of the control volume depends on the geometry and the flow conditions in the reactor. For example, for a plug flow reactor, the control volume is a cylinder of thickness dz (example of Fig. 1.9), for a tubular reactor with both radial and axial variations in process variables, the control volume will be a ring with height dz and radial thickness

dr, and for a tubular reactor with variations of process variables in radial, axial and angular directions, the control volume will be a sector of a ring.

Consider a cylindrical control volume between z and $z + dz$ at time t. All the variables such as the density ρ, velocity v and concentration C_A depend on time t and axial dimension z. The mass balance is performed on the control volume with thickness dz, and cross-sectional area S, giving

$$v(z)\, S\, \rho(z) = v(z + dz)\, S\, \rho(z + dz) + \frac{\partial(S\, dz\, \rho(z))}{\partial t} \tag{1.39}$$

which is simplified to

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho\, v)}{\partial z} = 0 \tag{1.40}$$

Diffusion is related to the axial concentration gradient in the reactor.

The axial diffusive flux of component A (moles per unit time and unit cross-sectional area) is expressed by Fick's law

$$N_A = -\mathcal{D}_A \frac{\partial C_A}{\partial z} \tag{1.41}$$

where $\mathcal{D}_A$ is the turbulent diffusion coefficient and the corresponding mass balance on A over the control volume is

$$\begin{aligned} v(z)\, S\, C_A(z) + S\, N_A(z) = \quad & v(z + dz)\, S\, C_A(z + dz) + S\, N_A(z + dz) + \\ & k\, C_A\ S\, dz + \frac{\partial(S\, dz\, C_A)}{\partial t} \end{aligned} \tag{1.42}$$

which can be simplified as

$$\frac{\partial C_A}{\partial t} + \frac{\partial v\, C_A}{\partial z} + k\, C_A - \frac{\partial}{\partial z}\left(\mathcal{D}_A \frac{\partial C_A}{\partial z}\right) = 0 \tag{1.43}$$

Such equations, however, are too complex for control applications. It is well known that a tubular reactor can be represented as a series of n perfectly mixed continuous stirred tank reactors, where n is large (infinite in theory). For control purposes, an approximate dynamic model is often sufficient. For example, a tubular reactor can be modelled as a cascade of a few (possibly three) perfectly mixed continuous stirred tank reactors, i.e. for i $(1 \leq i \leq 3)$ the overall and component mass balances become

$$\begin{cases} \rho_{i-1}\, v_{i-1} = & \rho_i\, v_i + \Delta z\, \dfrac{d\rho_i}{dt} \\ v_{i-1}\, C_{A,i-1} = & v_i\, C_{A,i} + k/S\, C_{A,i} + \Delta z\, \dfrac{dC_{A,i}}{dt} \end{cases} \tag{1.44}$$

Note that a rigorous simulation model will need a much larger number of elementary reactors in series.

1.6.3 Degrees of Freedom

A state-space model can represent either the steady state or the transient behaviour of a system. The steady-state solution obtained by setting all time derivatives to zero constitutes the intialization of the dynamic regime. The following discussion on degrees of freedom could be applied to steady state, but is here devoted to control, thus to the dynamic model.

The number of degrees of freedom *ndf* of a system is equal to the number of variables minus the number of equations

$$ndf = \text{number of var.} - \text{number of eq.}$$

If the degrees of freedom *ndf* is zero, the system is fully determined (or specified) and there exists only a unique solution; if it is positive, the system is under-specified and *ndf* equations should be added; if it is negative, the system is over-specified and *ndf* equations should be removed to get a unique solution.

Each control loop adds an additional equation. Furthermore, the external disturbances are also specified to reduce the number of unknowns.

1.7 Process Stability

A process is said to be stable (asymptotically) when in response to a disturbance, the state variables converge towards a steady state (the system is said to be feedback-negative). Another definition of stability is that a process is said to be stable if any bounded input results in a bounded output. If a bounded input results in an unbounded output, the process is unstable.

The process is unstable when in response to a disturbance some state variables tend mathematically towards infinity (the system is feedback-positive). In practice, that means simply that the variables go out of the desired domain or do not tend to come back in a stable manner, but oppositely go far from it at least periodically.

Nearly all chemical processes are stable in open loop. However, a CSTR with an exothermic reaction can be unstable. Indeed, if the cooling is insufficient with regard to the heat of reaction, there may be three stationary states, one stable at low temperature and low conversion, one stable at high temperature and high conversion, a third unstable state at an intermediate temperature and conversion (see Sect. 3.2.3). Nearly all processes can be made unstable in closed loop. A major recommendation for the controller design is to avoid instability.

1.8 Order of a System

If a process is described by an ordinary differential equation of order n, the process is said to be of order n

$$f(t) = a_0\, x + a_1 \frac{dx}{dt} + a_2 \frac{d^2x}{dt^2} + \ldots + a_n \frac{d^nx}{dt^n} \tag{1.45}$$

where $f(t)$ represents an input or a forcing function. Note that an ordinary differential equation of order n is equivalent to a set of n first-order ordinary differential equations.

The first-order process model

$$f(t) = a_0\, x + a_1\, \frac{dx}{dt} \tag{1.46}$$

can be written as

$$f(t) = a_0\, (x + \tau\, \frac{dx}{dt}) \tag{1.47}$$

where τ is the process time constant.

The second-order process model

$$f(t) = a_0\, x + a_1\, \frac{dx}{dt} + a_2\, \frac{d^2x}{dt^2} \tag{1.48}$$

can be written as

$$f(t) = a_0\, (x + 2\,\zeta\,\tau\, \frac{dx}{dt} + \tau^2\, \frac{d^2x}{dt^2}) \tag{1.49}$$

where ζ is the damping coefficient.

1.9 Laplace Transform

The Laplace transform is an elegant mathematical method to solve *linear* or *linearized* differential equations. In control theory, it is used to develop simple continuous input-output models and thereby analyze the influence of external variables on a given process.

The Laplace transform of a function $f(t)$ is defined by

$$\mathcal{L}[f(t)] = F(s) = \int_0^\infty f(t)\ \exp(-st)\, dt \tag{1.50}$$

assuming that the function or signal $f(t)$ is zero for $t < 0$. This is the monolateral Laplace transform which is used throughout this book. Notice that the exponential term has no dimension, therefore the dimension of variable s is the inverse of time (frequency).

If function $f(t)$ presents discontinuities at the boundaries, the Laplace transform is defined as

$$\mathcal{L}[f(t)] = F(s) = \lim \int_\epsilon^T f(t)\ \exp(-st)\, dt \qquad (\epsilon \to 0\ ,\ T \to \infty) \tag{1.51}$$

The bilateral Laplace transform for nonzero functions for negative t is defined as

$$\mathcal{L}[f(t)] = F(s) = \int_{-\infty}^\infty f(t)\ \exp(-st)\, dt \tag{1.52}$$

and is identical to the Fourier transform if we set $s = j\,\omega$. The Laplace transform exists only if the integral (1.50) is bounded: the function $f(t)\exp(-st)$ is

summable in Lebesgue's way. For example, consider the function $f(t) = \exp(t)$. This function is unbounded when $t \to +\infty$. However, let us try to calculate its Laplace transform. We get

$$\begin{aligned} \mathcal{L}[f(t)] &= \int_0^{\infty} \exp(t)\, \exp(-st)\, dt = \left[\frac{1}{s-1}\, \mathrm{e}^{(1-s)t}\right]_0^{+\infty} \\ &= \frac{1}{1-s} \qquad \text{if: } \mathcal{R}e(s) > 1 \end{aligned} \tag{1.53}$$

Its Laplace transform would be defined only in a frequency domain excluding low frequencies. Its convergence region is the domain of s complex values such that the Laplace transform exists, here the real part of s should be larger than 1.

If we consider the step function $f(t) = 1$, if $t \geq 0$, $f(t) = 0$ else, its Laplace transform is

$$\begin{aligned} \mathcal{L}[f(t)] &= \int_0^{\infty} 1\, \exp(-st)\, dt = \left[\mathrm{e}^{(-s)t}\right]_0^{+\infty} \\ &= \frac{1}{s} \qquad \text{if: } \mathcal{R}e(s) > 0 \end{aligned} \tag{1.54}$$

Its convergence region is the complex right half-plane.

The inverse Laplace transform is defined as

$$f(t) = \mathcal{L}^{-1}[F(s)] = \frac{1}{2\,\pi\, j} \int_{\sigma - j\,\infty}^{\sigma + j\,\infty} F(s)\, \exp(ts)\, ds \tag{1.55}$$

This integral is defined on a complex domain with $s = \sigma + j\,\omega$.

The Laplace transformation is a linear mapping

$$\mathcal{L}[a_1\, f_1(t) + a_2\, f_2(t)] = a_1\, \mathcal{L}[f_1(t)] + a_2\, \mathcal{L}[f_2(t)] \tag{1.56}$$

To get the inverse Laplace transform of $F(s)$, it is in general useful to expand the function $F(s)$ which very often takes the form of a rational fraction, as a sum of simple rational fractions, and then to operate the inverse transformation on each fraction separately.

1.9.1 Linearization and Deviation Variables

What is a Nonlinear Model?

Model linearization often poses problems for students. One reason may be that the analysis of the nonlinearity of the model is not clear. First, let us give some mathematical examples.

Consider a function of a single variable $f(x)$:
$f(x) = 4\,x$ and $f(x) = 2\,x + 3$ are linear with respect to x,
$f(x) = 3\,x^2$, $f(x) = \frac{1}{2\,x+3}$, $f(x) = \sqrt{3\,x}$, are nonlinear with respect to x.

A function of a single variable is linear with respect to this variable when the derivative of this function is constant. Otherwise, it is nonlinear.

Consider a function of two variables $f(x, y)$:
$f(x, y) = 2\,y + 3$ is linear with respect to y and independent of x,
$f(x, y) = 2\,x + 6\,y + 5$ is linear with respect to x and y,
$f(x, y) = x\,y$ is nonlinear with respect to x and y,
$f(x, y) = 2\,x^2 + 3\,y$ is nonlinear with respect to x and linear with respect to y,
$f(x, y) = 4\;x + 2\,x^2 + 3\,y + 5\,y^2$ is nonlinear with respect to x and y.

A function of several variables is linear with respect to one of its variables when the partial derivative of this function with respect to the considered variable is constant. Otherwise, it is nonlinear with respect to that variable.

Significance of Linearization in Process Control

Two cases can be considered: either a fixed set point is imposed on the process (case of regulation), or a reference trajectory is to be followed by the process (case of output tracking). Consider the simpler first case. We wish the process to be maintained in the neighbourhood of a set point, which will thus be the steady state. Due to imperfections of the control, the output and the state variables move around their steady-state values. The difference between the transient value of any variable and its steady-state value is called a deviation variable. When this latter remains small with respect to its absolute value, the behaviour of a function of this variable can be approximated by the tangent at the considered point (Fig. 1.10)

$$f(x) \approx f(x^s) + f'(x^s)\,(x - x^s) \tag{1.57}$$

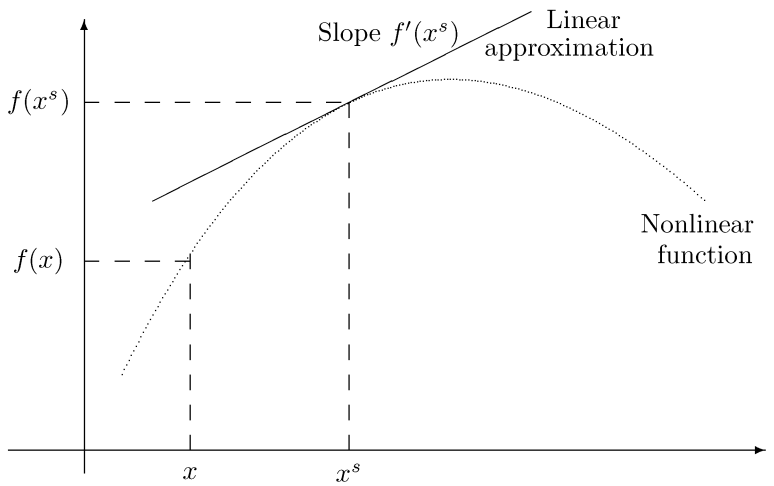


Figure 1.10: Linear approximation of a nonlinear function by linearization of this function around steady state

General Linearization

A nonlinear state-space model of a process must be linearized before applying the Laplace transform. The system thus obtained is called a linear time-invariant (LTI) system. Consider again the example of the chemical reactor (Eq. 1.16); the term FC_A, which appears in the differential equation in the component mass balance, is nonlinear as $F(t)$ and $C_A(t)$ both depend on time. As the Laplace transformation is a linear mapping, it is therefore necessary to linearize the balance ordinary differential equations around steady state (denoted by ^{s}); thus, the product $F\,C_A$ becomes

$$F(t)\,C_A(t) = F^s\,C_A^s + F^s\,(C_A(t) - C_A^s) + C_A^s\,(F(t) - F^s) + 0(\epsilon^2) \tag{1.58}$$

The term $0(\epsilon^2)$ indicates that the Taylor series expansion in the neighbourhood of the steady state is truncated at the first order. Generally, for a function f of n variables $x_1, \ldots, x_n$, by neglecting the higher-order terms, the Taylor series expansion leads to

$$f(x_1, \ldots, x_n) \approx f(x_1^s, \ldots, x_n^s) + \sum_i \left(\frac{\partial f}{\partial x_i}\right)^s (x_i - x_i^s) \Longleftrightarrow \delta f \approx \sum_i \left(\frac{\partial f}{\partial x_i}\right)^s \delta x_i \tag{1.59}$$

It can be noticed that the linearization results in deviation variables (with respect to the steady state or a reference state) of the form $\delta x_i = (x_i - x_i^s)$ which play an important role in Laplace transformation.

1.9.2 Some Important Properties of Laplace Transformation

- The Laplace transformation is a linear operation

$$\mathcal{L}[a_1\,f_1(t) + a_2\,f_2(t)] = a_1\,\mathcal{L}[f_1(t)] + a_2\,\mathcal{L}[f_2(t)] \tag{1.60}$$

- The Laplace transform of a first-order derivative of a function is

$$\mathcal{L}[\frac{df(t)}{dt}] = s\,F(s) - f(0) \tag{1.61}$$

If $f(t)$ is a deviation variable with respect to the initial steady state, its initial value becomes zero: $f(0) = 0$, and the previous equation simply becomes

$$\mathcal{L}[\frac{df(t)}{dt}] = s\,F(s) \tag{1.62}$$

This assumption is used in general.

- The Laplace transform of the n^{th}-order derivative of a function is

$$\mathcal{L}[\frac{d^n f(t)}{dt^n}] = s^n\,F(s) - s^{n-1}\,f(0) - s^{n-2}\,f^{(1)}(0) \ldots - f^{(n-1)}(0) \tag{1.63}$$

If $f(t)$ is a deviation variable, its initial value and successive derivatives up to the $(n-1)$-th-order become zero so that the previous formula becomes

$$\mathcal{L}[\frac{d^n f(t)}{dt^n}] = s^n\, F(s) \tag{1.64}$$

• The Laplace transform of the integral of a function is

$$\mathcal{L}[\int_0^t f(x)\, dx] = \frac{1}{s}\, F(s) \tag{1.65}$$

• The initial value theorem is

$$\lim_{t \to 0} f(t) = \lim_{s \to \infty} s\, F(s) \tag{1.66}$$

• The final value theorem is

$$\lim_{t \to +\infty} f(t) = \lim_{s \to 0} s\, F(s) \tag{1.67}$$

Notice that the final value theorem cannot be applied in the case of a function corresponding to an unstable system. For example, consider the Laplace transform $F(s) = 1/(s-1)$, which would be the transform of function $f(t) = \exp(t)$ if we strictly apply Table 1.1. Let us try to apply the final value theorem (1.67) to this function. It gives

$$\lim_{t \to +\infty} f(t) = \lim_{s \to 0} \frac{s}{s-1} = 0^- \tag{1.68}$$

One would wrongly conclude that the function $f(t) = \exp(t)$ tends towards 0 when $t \to +\infty$. The mistakes comes from the fact that one does not take into account the remark concerning Eq. (1.53).

Consider the Laplace transform $F(s) = 1/(s(s+1))$ corresponding to a stable system (we will later see that it is a first-order system subjected to an input step). The final value theorem gives

$$\lim_{t \to +\infty} f(t) = \lim_{s \to 0} \frac{s}{s(s+1)} = 1 \tag{1.69}$$

• The Laplace transform of a delayed function is

$$\mathcal{L}[f(t-t_0)] = \exp(-s\, t_0)\, \mathcal{L}[f(t)] \tag{1.70}$$

Note that the function $f(t-t_0)$ is the same as $f(t)$ delayed by t_0, which means that at time t_0, the delayed function is equal to $f(0)$ (Fig. 1.11). The delay corresponds to a time translation of the function.

• The complex translation is

$$\mathcal{L}[f(t)\, \exp(at)] = F(s-a) \tag{1.71}$$

• The scale change is

$$\mathcal{L}[f(t/a)] = a\, F(a\, s) \tag{1.72}$$

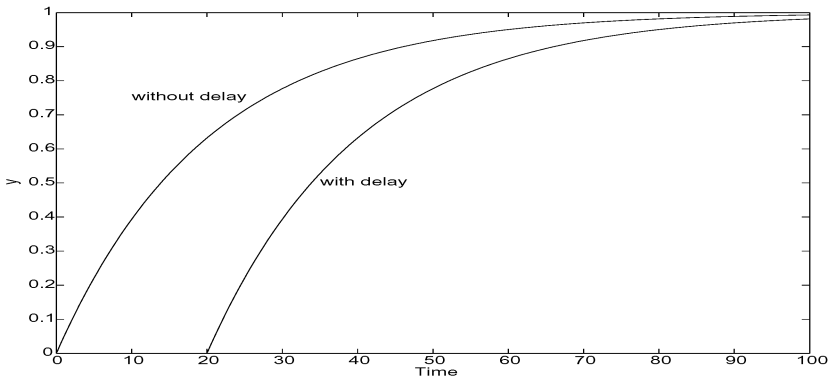


Figure 1.11: Illustration of the effect of a time delay of 20 time units on a response

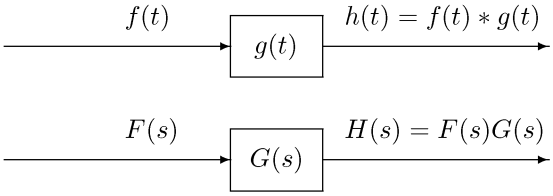


Figure 1.12: Convolution for a linear system. Top: time domain. Bottom: frequency domain

• Laplace transform of convolution
When a signal $f(t)$ excites a linear time-invariant system with impulse response $g(t)$ (Fig. 1.12), the response of the system $h(t)$ is equal to the convolution product of $f(t)$ by $g(t)$ denoted by

$$h(t) = f(t) * g(t) = \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau \tag{1.73}$$

The Laplace transform $H(s)$ of the output is equal to

$$\mathcal{L}[f(t) * g(t)] = F(s)\, G(s) \tag{1.74}$$

which is expressed as the Laplace transform of a convolution product being equal to the product of the Laplace transforms of the functions.

A consequence of this property is that when a signal $f(t)$ excites two linear systems in series $g_1(t)$ and $g_2(t)$, the response $h(t)$ is equal to

$$h(t) = f(t) * (g_1(t) * g_2(t)) \tag{1.75}$$

and its Laplace transform $H(s)$ is

$$\mathcal{L}[f(t) * (g_1(t) * g_2(t))] = F(s)\, G_1(s)\, G_2(s) \tag{1.76}$$

Thus, the Laplace transform of the impulse response of two linear systems in series is equal to the product of the Laplace transforms of the individual impulse responses.

• The complex convolution is

$$\mathcal{L}[f(t)\,g(t)] = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} F(q)\,G(s-q)\,dq \tag{1.77}$$

• Parseval-Plancherel relation

This relation, classical in signal processing, expresses that the energy of a signal is equal to the sum of the energies of its constitutive signals

$$\int_{-\infty}^{+\infty} f^2(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |F(j\omega)|^2\,d\omega \tag{1.78}$$

It is necessary that the signal $f(t)$ be square-integrable, which means that the integral of $f^2(t)$ must exist.

• Differentiation or integration with respect to a parameter

Consider the function $f(t,\theta)$ which depends on parameter θ as well as time t (both are independent variables) the Laplace transforms of the derivative or integral of the function with respect to θ are

$$\mathcal{L}\left[\frac{\partial f(t,\theta)}{\partial\theta}\right] = \frac{\partial}{\partial\theta}\mathcal{L}[f(t,\theta)] \tag{1.79}$$

and

$$\mathcal{L}\left[\int_{\theta_1}^{\theta_2} f(t,\theta)\,d\theta\right] = \int_{\theta_1}^{\theta_2} \mathcal{L}[f(t,\theta)]\,d\theta \tag{1.80}$$

• Table 1.1 lists the Laplace transforms of some common functions.

1.9.3 Transfer Function

Definition of a Transfer Function

Consider a linear single variable system whose dynamic behaviour is described in terms of deviation variables by an ordinary differential equation of order n linking the input and the output

$$b_0\,u(t)+b_1\frac{du(t)}{dt}+\ldots+b_m\frac{d^m u(t)}{dt^m} = a_0\,y(t)+a_1\frac{dy(t)}{dt}+\ldots+a_n\frac{d^n y(t)}{dt^n} \quad , \; m \le n \tag{1.81}$$

where $u(t)$ and $y(t)$ are the system input and output respectively (Fig. 1.13). If we assume that the system is initially at steady state, the deviation variables and their successive derivatives are zero at initial state

$$\begin{aligned} \delta u(0) = 0, \left(\frac{du}{dt}\right)_0 = 0, \ldots, \left(\frac{d^m u}{dt^m}\right)_0 = 0 \\ \delta y(0) = 0, \left(\frac{dy}{dt}\right)_0 = 0, \ldots, \left(\frac{d^n y}{dt^n}\right)_0 = 0 \end{aligned} \tag{1.82}$$

Table 1.1: Laplace transform of some common functions

Signal $f(t) \quad (t \geq 0)$	Transform $\mathcal{L}[f(t)] = F(s)$
Convolution product $f(t) * g(t)$	$F(s)\,G(s)$
Derivative: $\frac{df(t)}{dt}$	$s\,F(s) - f(0)$
Integral: $\int_0^t f(x)\,dx$	$\frac{1}{s}\,F(s)$
Delayed function: $f(t - t_0)$	$\exp(-s\,t_0)\,F(s)$
Dirac unit impulse: $\delta(t)$	1
Unit impulse of duration τ defined by $\delta_\tau(t) = 0$ if $t < 0$ or $t > \tau$ $\delta_\tau(t) = \frac{1}{\tau}$ if $0 < t < \tau$	$\frac{1}{\tau}\,\frac{1 - \exp(-s/\tau)}{s}$
Step of amplitude A	$\frac{A}{s}$
Exponential: $\exp(-a\,t)$	$\frac{1}{s+a}$
$\frac{1}{\tau}\,\exp(-t/\tau)$	$\frac{1}{\tau\,s+1}$
Ramp: $a\,t$	$\frac{a}{s^2}$
$t\,\exp(-at)$	$\frac{1}{(s+a)^2}$
$\frac{1}{n!}\,t^n\,\exp(-at)\;(n \geq 1)$	$\frac{1}{(s+a)^{n+1}}$
$\sin(\omega\,t)$	$\frac{\omega}{s^2+\omega^2}$
$\cos(\omega\,t)$	$\frac{s}{s^2+\omega^2}$
$\sin(\omega\,t+\phi)$	$\frac{\omega\,\cos(\phi) + s\,\sin(\phi)}{s^2+\omega^2}$
$\exp(-at)\,\sin(\omega t)$	$\frac{\omega}{(s+a)^2+\omega^2}$
$\exp(-at)\,\cos(\omega t)$	$\frac{s+a}{(s+a)^2+\omega^2}$
$\frac{t^n}{n!}$	$\frac{1}{s^{n+1}}$
$\frac{1}{b-a}\,(\exp(-at) - \exp(-bt))$	$\frac{1}{(s+a)\,(s+b)}$
$\frac{1}{\tau_1-\tau_2}\,(\exp(-t/\tau_1) - \exp(-t/\tau_2))$	$\frac{1}{(\tau_1 s+1)\,(\tau_2 s+b)}$
$\sum_{i=1}^{n}\left(\frac{\exp(-a_i\,t)}{\prod_{j\neq i}(a_j - a_i)}\right)$	$\frac{1}{\prod_{i=1}^{n}(s+a_i)}$
$\frac{1}{\omega_p}\,\exp(-\zeta\,\omega\,t)\,\sin(\omega_p\,t)$ with: $\omega_p = \omega\,\sqrt{1-\zeta^2} \quad (0 \leq \lvert\zeta\rvert \leq 1)$	$\frac{1}{s^2 + 2\,\zeta\omega\,s + \omega^2}$
$\frac{1}{a}\,(1 - \exp(-a\,t))$	$\frac{1}{s(s+a)}$

The Laplace transformation of (1.81) gives

$$(b_0 + b_1\,s + \ldots + b_m\,s^m)\,U(s) = (a_0 + a_1\,s + \ldots + a_n\,s^n)\,Y(s) \tag{1.83}$$

The transfer function of the system results as the ratio of the Laplace trans-

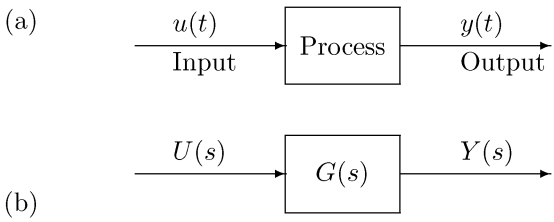


Figure 1.13: Block diagram of a process (a) in the time domain, (b) in the Laplace domain (transfer function)

form of the output variable to the Laplace transform of the input variable, both expressed in terms of deviations from their steady states

$$\frac{Y(s)}{U(s)} = G(s) = \frac{b_0 + b_1\, s + \ldots + b_m\, s^m}{a_0 + a_1\, s + \ldots + a_n\, s^n} \tag{1.84}$$

Thus, the transfer function is totally equivalent to the linear ordinary differential equation describing the linearized system and can be further used to find output solutions to a given input.
Most transfer functions take the previous form of a ratio of two polynomials, symbolized as

$$G(s) = \frac{N(s)}{D(s)} \tag{1.85}$$

A transfer function is said to be proper if

$$\text{degree of } N(s) \leq \text{degree of } D(s)$$

it is strictly proper if

$$\text{degree of } N(s) < \text{degree of } D(s)$$

A transfer function is said to be biproper if

$$\text{degree of } N(s) = \text{degree of } D(s)$$

A transfer function is said to be improper if

$$\text{degree of } N(s) > \text{degree of } D(s)$$

It will be shown that improper transfer functions, such as an ideal derivative, amplify high-frequency noise.

The following steps are followed to **derive the transfer function of a process from a theoretical model**:

- Using the conservation principles, write the dynamic model describing the system,
- Linearize equations using Taylor series expansion,
- Express the equations in terms of deviation variables by subtracting the steady-state equations from the dynamic equations
- Operate Laplace transformation on the linear or linearized equations,
- Obtain the ratio of the Laplace transform of the output over the Laplace transform of the input.

Example 1.1: **Application to the Surge Tank**

Equation (1.13) is linear with respect to all variables, and results simply in

$$X(s) = \frac{1}{S\,s}\,U(s) - \frac{1}{S\,s}\,\bar{F}(s) \tag{1.86}$$

which contains two transfer functions, both pure integrators. The process transfer function with respect to the input is

$$G_u(s) = \frac{1}{S\,s} \tag{1.87}$$

and the load or disturbance transfer function with respect to the disturbance F is

$$G_d(s) = -\frac{1}{S\,s} \tag{1.88}$$

where S is the cross-sectional area of the surge tank and s is the Laplace operator.

The block diagram (Fig. 1.14) represents the influence of the input and of the disturbance.

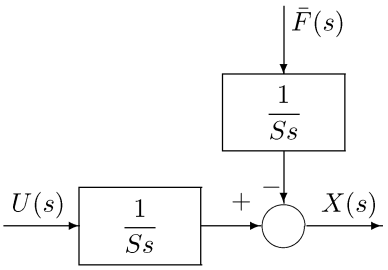


Figure 1.14: Block diagram of the surge tank

Example 1.2: **Application to the Isothermal Chemical Reactor**

The system of Eqs. (1.27) is nonlinear with respect to the variables. Using the superscript "s" for the steady-state, linearizing the nonlinear terms, introducing the equations in terms of deviation variables and Laplace transforming results in

$$\begin{aligned}
sX_1(s) = \;& \frac{1}{S}[\bar{F}_0(s) - \bar{F}(s) + U_1(s)] \\
sX_2(s) = \;& -\frac{1}{S(x_1^s)^2}\,[F_0^s\,(u_2^s - x_2^s)X_1(s)] - k\,X_2(s) \\
& + \frac{1}{Sx_1^s}\,\left[(u_2^s - x_2^s)\bar{F}_0(s) + F_0^s\,(U_2(s) - X_2(s)) - x_2^s\,U_1(s) - u_1^s\,X_2(s)\right]
\end{aligned} \tag{1.89}$$

The steady-state equations resulting from Eq. (1.27) are

$$\begin{aligned} 0 = &\quad \frac{1}{S}[F_0^s - F^s + u_1^s] \\ 0 = &\quad \frac{1}{S\,x_1^s}\,[F_0^s\,(u_2^s - x_2^s) - x_2^s\,u_1^s] - k\,x_2^s \end{aligned} \tag{1.90}$$

which yield

$$\begin{aligned} u_1^s = &\quad F^s - F_0^s \\ u_2^s = &\quad \frac{x_2^s\,F^s + S\,k\,x_1^s\,x_2^s}{F_0^s} \end{aligned} \tag{1.91}$$

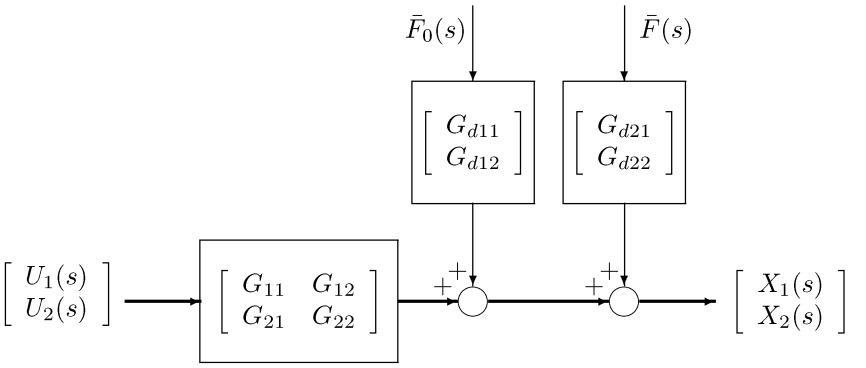


Figure 1.15: Block diagram of the linearized isothermal chemical reactor tank

The final system of transfer functions, the transfer function matrix, expresses the state vector with respect to the input vector and the disturbances

$$\begin{aligned} \begin{bmatrix} X_1(s) \\ X_2(s) \end{bmatrix} = &\begin{bmatrix} \dfrac{\dfrac{1}{S\,s}}{s + \dfrac{F^s}{S x_1^s} + k} & 0 \\ \dfrac{-\dfrac{1}{Ss}\dfrac{x_2^s}{S(x_1^s)^2}(F^s + Skx_1^s - F_0^s) - \dfrac{x_2^s}{Sx_1^s}}{s + \dfrac{F^s}{Sx_1^s} + k} & \dfrac{\dfrac{F_0^s}{Sx_1^s}}{s + \dfrac{F^s}{Sx_1^s} + k} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} \\ &+ \begin{bmatrix} \dfrac{\dfrac{1}{S\,s}}{s + \dfrac{F^s}{Sx_1^s} + k} \\ \dfrac{\dfrac{x_2^s}{Sx_1^s}(\dfrac{1}{F_0^s} - \dfrac{1}{Ssx_1^s})(F^s + Skx_1^s - F_0^s)}{s + \dfrac{F^s}{Sx_1^s} + k} \end{bmatrix} \bar{F}_0(s) \\ &+ \begin{bmatrix} \dfrac{-\dfrac{1}{S\,s}}{s + \dfrac{F^s}{Sx_1^s} + k} \\ \dfrac{1}{Ss}\dfrac{x_2^s}{S(x_1^s)^2}\dfrac{F^s + Skx_1^s - F_0^s}{s + \dfrac{F^s}{Sx_1^s} + k} \end{bmatrix} \bar{F}(s) \end{aligned} \tag{1.92}$$

Note that the states x_1 and x_2 are coupled by this system of equations. The first state x_1 is influenced only by input u_1, while the second state x_2 is influenced by both inputs. With the system being multi-input multi-state, matrices of transfer functions relate the inputs and disturbances to the states. The linearized system described by the previous equation can be symbolized by the block diagram in Fig. 1.15.

The non-isothermal chemical reactor could be treated in the same manner with added complexity.

Impulse Response and Transfer Function of a System

Consider a linear system in which the input and output are linked by the convolution product (Fig. 1.16)

$$y(t) = u(t) * g(t) \tag{1.93}$$

If this system is excited by a Dirac impulse: $u(t) = \delta(t)$, the output becomes

$$y(t) = g(t) \tag{1.94}$$

thus $g(t)$ is the impulse response of the system and is equal to the inverse Laplace transform of the system transfer function as the Laplace transform of the convolution product is

$$\mathcal{L}[y(t)] = \mathcal{L}[u(t) * g(t)] = U(s)\,G(s) = G(s) \tag{1.95}$$

with $U(s) = 1$ for a unit impulse input.

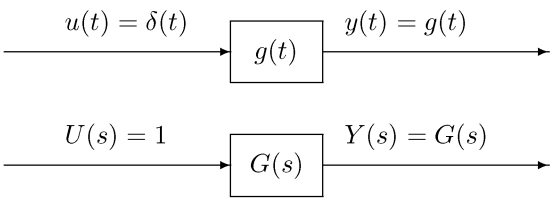


Figure 1.16: Time and frequency responses for a linear system subjected to a Dirac impulse input

Although obtaining a system transfer function in this manner seems attractive, it suffers from several drawbacks:

- A Dirac impulse is not realizable in practice, only an impulse of finite duration is possible.
- A simple impulse input contains poor characteristics with respect to frequency excitation and introduces difficulty in process identification as will be shown later in this book.

In experimental characterization of the flow pattern in chemical processes, in particular reactors, however, the impulse response technique is used to get the residence time distribution. A tracer is injected over a short time at the process inlet and its evolution is recorded at the reactor outlet. The analysis of the output response is often complicated but it enables us to characterize the system as a combination of perfectly mixed CSTRs, tubular reactors placed in parallel or in series with possible bypass and/or dead volume (Levenspiel, 1999; Villermaux, 1982).

Principle of superposition

When the input can be decomposed into a sum of inputs (e.g. a rectangular pulse is the sum of a positive and a negative step, occurring at different times, see Figs. 1.17 and 1.18), the global output is the sum of the responses to the individual inputs

$$U(s) = \sum_i U_i(s) \implies Y(s) = G(s)\,U(s) = \sum_i G(s)\,U_i(s) = \sum_i Y_i(s) \quad (1.96)$$

The principle of superposition results directly from the properties of linearity of the Laplace transform. The principle of superposition is not valid for a nonlinear system.

For example, a rectangular pulse can be seen as the sum of a positive and a negative step (in Fig. 1.17, u_3 is the sum of u_1 and u_2). The respective responses to the three inputs for a given transfer function are given in Fig. 1.18.

Experimental Determination (Identification) of a System Transfer Function

Consider a general system whose dynamics in the time domain can be described in terms of the following ordinary differential equation

$$\frac{dy}{dt} = f(y(t)\,,u(t)\,,d\,) \quad (1.97)$$

where y is the output, u is the input, and d is the disturbance. To determine the system transfer function $G(s)$, several possibilities exist depending on the choice of the applied input to the system:

- For a Dirac impulse input, the output is equal to $g(t)$, i.e. the inverse transform of the system transfer function $G(s)$. The drawback of this technique is that the frequency content (the information content) of the input signal is not rich and results in poor identification.
- For a step input or a succession of small amplitude positive and negative steps such as a pseudo-random binary sequence (PRBS), the information content of the input signal is adequate for yielding a satisfactory identification. This technique is particularly well adapted to discrete-time identification. In the continuous-time domain, often a single step or a succession

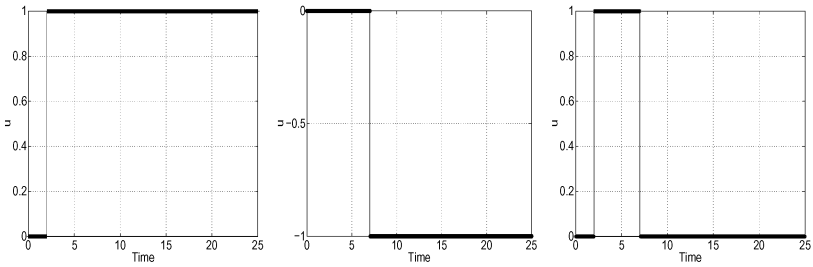


Figure 1.17: Inputs from left to right: u_1, u_2, u_3

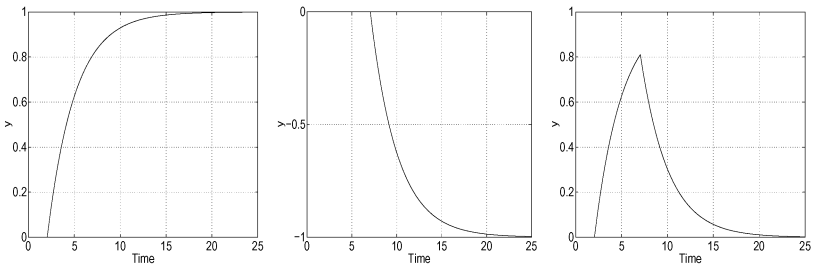


Figure 1.18: Outputs from left to right: y_1, y_2, y_3 as responses of the transfer function $G(s) = \dfrac{1}{3s+1}$ to respective inputs u_1, u_2, u_3

of positive and negative steps is applied. System nonlinearities are demonstrated by dissimilar responses to positive and negative step changes in the input. For example, a furnace often shows a different time constant for a step increase or decrease in the heat rate.

– For a sinusoidal input (Fig. 1.19), the identification is performed in the frequency domain. In this latter method, the Laplace variable s is replaced by $s = j\,\omega$ (where ω is the angular frequency in rad/s and is related to the frequency in Hz, ν, by $\omega = 2\,\pi\,\nu$). If a linear system is excited by a sinusoidal input $u(t) = \exp(j\,\omega\,t)$, after a transient period, the output will also be a sinusoidal wave with the same frequency, a different amplitude and a phase difference, i.e.

$$y(t) \approx G(j\,\omega\,)\,\exp(j\,\omega\,t) = G(s)\,\exp(st) \qquad \text{for sufficiently large } t \quad (1.98)$$

which, when combined with Eq. (1.97), gives

$$s\,G(s)\,\exp(st) = f(G(s)\exp(st),\,\exp(st),\,d) \quad (1.99)$$

Complex exponential functions (such as sines and cosines) are proper functions for the Laplace operator. In order to obtain a rich information for the experimental determination of the transfer functions, it is sufficient to vary the frequency ω of the input signal over a large range.

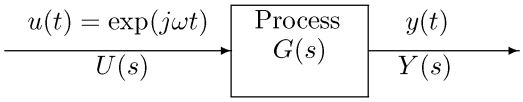


Figure 1.19: Experimental determination of a process transfer function by sinusoidal excitation

Provided the function f is known, one analytically deduces from Eq. (1.99) the transfer function $G(s)$ as the ratio of the output transform $Y(s)$ over the input transform $U(s)$.

1.9.4 Poles and Zeros of a Transfer Function

Consider a system described by the differential equation

$$a_0\, y(t) + a_1 \frac{dy(t)}{dt} + \ldots + a_n \frac{d^n y(t)}{dt^n} = b_0\, u(t) + b_1 \frac{du(t)}{dt} + \ldots + b_m \frac{d^m u(t)}{dt^m} \quad (1.100)$$

In the absence of an input excitation, the output Laplace transform is given by

$$D(s)\, Y(s) = 0 \quad (1.101)$$

where $D(s) = a_0 + a_1\, s + \ldots + a_n\, s^n$. The shape of the response curve depends on the roots of $D(s)$ which are called the system modes.

The system transfer function describes the response when the initial state is zero (refer to the relation between the transfer function and the state-space model) or when we refer to deviation variables. In general, the transfer function can be expressed as the ratio of two polynomials

$$G(s) = \frac{N(s)}{D(s)} = \frac{b_0 + b_1\, s + \ldots + b_m\, s^m}{a_0 + a_1\, s + \ldots + a_n\, s^n} \qquad \text{with:} \quad n \geq m \quad (1.102)$$

where the denominator degree n is higher than or equal to the numerator degree m.

The roots of the numerator polynomial $N(s)$ are called the system zeros or zeros of the transfer function, since the transfer function is equal to zero for these values.

If the numerator $N(s)$ and the denominator $D(s)$ have common roots, after cancellation of these common roots, the remaining roots of $D(s)$ are referred to as the poles of the transfer function; they determine the response for zero initial state.

If no common roots exist between $N(s)$ and $D(s)$, then the system can be completely characterized by its transfer function, and the sets of poles and the modes are the same. The transfer function $G(s)$ becomes infinite when s is equal to one of the poles.

If $N(s)$ and $D(s)$ have common factors, the latter form a polynomial noted $R(s)$. The roots of $R(s)$ are called system nodes, but are not poles for $G(s)$. For that reason, they are called missing poles of the transfer function. In this case, the system is not completely characterized by its transfer function.

1.9.5 Qualitative Analysis of a System Response

The response of a system to a given input $u(t)$ can be determined from its Laplace transform

$$Y(s) = G(s)\,U(s) \tag{1.103}$$

If the transfer function $G(s)$ and the input transform $U(s)$ are given, the system response $y(t)$ can be obtained by inverse Laplace transformation of $Y(s)$.

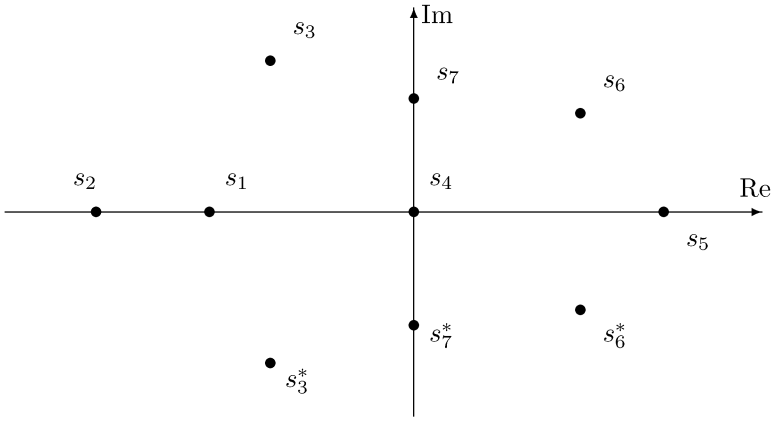


Figure 1.20: Different types of poles represented in the complex s-plane

The behaviour of a system, such as its stability and the the shape of the system response, is largely determined by the poles of its transfer function. Let us consider a general transfer function with poles as depicted in Fig. 1.20

$$G(s) = \frac{N(s)}{D(s)} = \frac{N(s)}{(s-s_1)\,(s-s_2)^m\,(s-s_3)\,(s-s_3^*)\,(s-s_4)\,(s-s_5)\,(s-s_6)\,(s-s_6^*)\,(s-s_7)\,(s-s_7^*)} \tag{1.104}$$

where s_1 is a negative real pole

s_2 is a negative real multiple pole of order m

s_3 is a complex pole with negative real part

s_3^* is the conjugate complex pole (D being a real polynomial, all complex poles must appear as a pair in conjugate form)

s_4 is a pole at the origin

s_5 is a positive real pole

s_6 is a complex pole with positive real part

s_6^* is the complex conjugate pole of s_6

s_7 is a pure imaginary pole

s_7^* is the conjugate of the imaginary pole s_7

The transfer function $G(s)$ can be expanded as a sum of rational fractions

$$G(s) = \frac{c_1}{s-s_1} + \left(\frac{c_{21}}{s-s_2} + \frac{c_{22}}{(s-s_2)^2} + \ldots + \frac{c_{2m}}{(s-s_2)^m} \right) + \frac{c_3}{s-s_3} + \frac{c_3^*}{s-s_3^*} + \frac{c_4}{s} + \frac{c_5}{s-s_5} + \frac{c_6}{s-s_6} + \frac{c_6^*}{s-s_6^*} + \frac{c_7}{s-s_7} + \frac{c_7^*}{s-s_7^*} \tag{1.105}$$

If poles are distinct, to find the coefficients in the numerators (residuals), it suffices to multiply $G(s)$ by $(s-s_i)$ and set $s = s_i$

$$c_i = \left[(s-s_i)\, G(s) \right]\big|_{s=s_i} = \left. \frac{N(s)}{(s-s_1)\ldots(s-s_{i-1})(s-s_{i+1})\ldots(s-s_n)} \right|_{s=s_i} \tag{1.106}$$

In the case of multiple poles, e.g. if s_i is a multiple pole of order m, the residuals are calculated by multiplying $G(s)$ by $(s-s_i)^m$ and differentiating successively with respect to s and setting $s = s_i$ after differentiation. Consider the following transfer function with a multiple pole s_i of order m

$$\begin{aligned} G(s) &= \frac{N(s)}{(s-s_1)\ldots(s-s_{i-1})\,(s-s_i)^m\,(s-s_{i+1})\ldots(s-s_n)} \\ &= \frac{c_1}{s-s_1} + \ldots + \frac{c_{i-1}}{s-s_{i-1}} + \frac{c_{i,1}}{s-s_i} + \frac{c_{i,2}}{(s-s_i)^2} + \ldots + \frac{c_{i,m}}{(s-s_i)^m} \\ &\quad + \frac{c_{i+1}}{s-s_{i+1}} + \ldots + \frac{c_n}{s-s_n} \end{aligned} \tag{1.107}$$

Let us define a new transfer function having the same poles as $G(s)$ except for the multiple pole s_i

$$Gt(s) = \frac{N(s)\,(s-s_i)^m}{D(s)} \tag{1.108}$$

The residuals are calculated by the following expression

$$c_{i,m-j} = \frac{1}{j!} \left. \frac{d^{(j)} Gt(s)}{ds^j} \right|_{s=s_i} \quad ; \quad j = 0, 1, \ldots, m-1 \tag{1.109}$$

Note that multiple poles are more frequently encountered in physical modelling than in identification.

When a process having a transfer function $G(s)$ of the form given by Eq. (1.107) is subjected to a unit impulse, the Laplace transform of the process output is equal to the process transfer function. To determine the time response to this impulse input, the inverse Laplace transformation is performed.

- Simple real poles such as s_1 and s_5 result in exponential responses (Fig. 1.21)

$$y(t) = c_i \exp(s_i\, t) \tag{1.110}$$

If the pole is negative real, the exponential tends towards 0: s_1 is a stable pole. The closer the pole to the origin, the slower the response will be. If the pole is positive real such as s_5, the response increases exponentially with time and tends towards infinity: s_5 is an unstable pole.

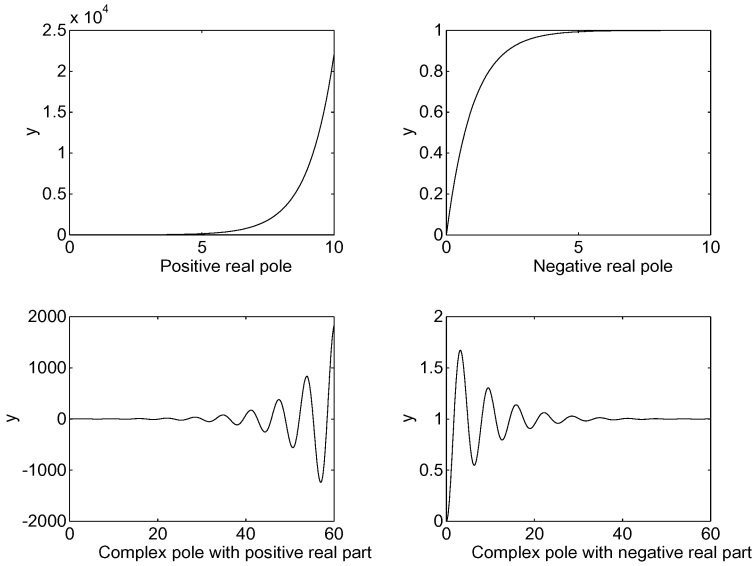


Figure 1.21: Impulse response of transfer functions with different types of poles

- A multiple real pole with order i such as s_2 results in the following responses:

$$y(t) = \frac{t^{i-1} \exp(s_2\, t)}{(i-1)!} \tag{1.111}$$

  The response increases towards infinity if the pole is positive or zero, and decreases towards zero if the pole is negative.
- A complex pole s is represented by its real part and its imaginary part: $s = s_R + j\, s_I$, and the impulse response of a transfer function presenting a complex pole will be

$$y(t) = \exp(s_R\, t)\, \exp(j\, s_I\, t) = \exp(s_R\, t)\, [\cos(s_I\, t) + j \sin(s_I\, t)] \tag{1.112}$$

  The imaginary part is responsible for the oscillatory behaviour while the stability depends only on the sign of the real part.
  Conjugate complex poles with negative real part such as s_3 and s_3^* (Fig. 1.21) result in damped oscillatory behaviour: they are stable poles. The closer the poles are to the imaginary axis, i.e. the nearest to 0 their real part is, the slower the decay of their exponential part.
  When poles are complex conjugate with positive real part such as s_6 and s_6^* (Fig. 1.21), they induce an oscillatory undamped response: they are unstable poles.
- A pole at the origin such as s_4 results in a constant term.

- Pure imaginary poles $s = \pm j\,\omega$ such as s_7 and s_7^* result in a sinusoidal response with a frequency ω equal to the imaginary part of the poles. Under these conditions the system is said to be on the verge of stability or marginally stable.

1.10 Linear Systems in State Space

1.10.1 General Case

Consider the single variable linear system (single-input single-output). Its state-space model is

$$\begin{cases} \dot{\boldsymbol{x}}(t) = & \boldsymbol{A}\,\boldsymbol{x}(t) + \boldsymbol{B}\,u(t) \\ y(t) = & \boldsymbol{C}\,\boldsymbol{x}(t) + D\,u(t) \end{cases} \tag{1.113}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, D are matrices of respective dimensions $n \times n$, $n \times 1$, $1 \times n$ and 1×1. The initial state of the system is $\boldsymbol{x}(0)$. For a system which can be represented by a strictly proper transfer function, i.e. whose numerator degree is strictly lower than the denominator degree, the matrix D is zero. This is the case for most physical systems.

As this system is linear, it is possible to apply the Laplace transformation

$$s\,X(s) - \boldsymbol{x}(0) = \boldsymbol{A}\,X(s) + \boldsymbol{B}\,U(s) \Longleftrightarrow (s\,\boldsymbol{I} - \boldsymbol{A})\,X(s) = \boldsymbol{x}(0) + \boldsymbol{B}\,U(s) \tag{1.114}$$

where $\boldsymbol{I}$ is the identity matrix of dimension $n \times n$. Provided that the matrix $(s\,\boldsymbol{I} - \boldsymbol{A})$ is invertible, the Laplace transform of the states can be obtained

$$X(s) = (s\,\boldsymbol{I} - \boldsymbol{A})^{-1}\,\boldsymbol{x}(0) + (s\,\boldsymbol{I} - \boldsymbol{A})^{-1}\,\boldsymbol{B}\,U(s) \tag{1.115}$$

and the Laplace transform of the output is

$$Y(s) = \boldsymbol{C}\,X(s) + D\,U(s) = \boldsymbol{C}\,(s\,\boldsymbol{I} - \boldsymbol{A})^{-1}\,\boldsymbol{x}(0) + \left[\boldsymbol{C}\,(s\,\boldsymbol{I} - \boldsymbol{A})^{-1}\,\boldsymbol{B} + D\right]\,U(s) \tag{1.116}$$

This response is composed of two terms, the first called response for a zero input and the second called response for a zero state.

Given

$$\exp(\boldsymbol{A}\,t) = \mathcal{L}^{-1}\left[(s\,\boldsymbol{I} - \boldsymbol{A})^{-1}\right] \tag{1.117}$$

the output response can be deduced from (1.116)

$$y(t) = \boldsymbol{C}\,\exp(\boldsymbol{A}\,t)\,\boldsymbol{x}(0) + \boldsymbol{C}\,\exp(\boldsymbol{A}\,t)\int_0^t \exp(-\boldsymbol{A}\,\tau)\,\boldsymbol{B}\,u(\tau)\,d\tau + D\,u(t) \tag{1.118}$$

and the state of the process is given by

$$x(t) = \exp(\boldsymbol{A}\,t)\,\boldsymbol{x}(0) + \exp(\boldsymbol{A}\,t)\int_0^t \exp(-\boldsymbol{A}\,\tau)\,\boldsymbol{B}\,u(\tau)\,d\tau \tag{1.119}$$

To obtain $y(t)$ as in (1.118), it is possible to integrate the differential equation of the system given by (1.113) by using the following properties of the matrix exponential

$$\exp(\boldsymbol{A}\,t) = \boldsymbol{I} + t\,\boldsymbol{A} + \frac{t^2}{2!}\,\boldsymbol{A}^2 + \ldots + \frac{t^n}{n!}\,\boldsymbol{A}^n + \ldots \tag{1.120}$$

$$\frac{d}{dt}(\exp(\boldsymbol{A}\,t)) = \boldsymbol{A}\,\exp(\boldsymbol{A}\,t) \tag{1.121}$$

The system transfer function is obtained, assuming that all initial conditions are zero, and consequently the initial states are zero: $\boldsymbol{x}(0) = 0$. The output Laplace transform is

$$Y(s) = \left[\boldsymbol{C}\,(s\,\boldsymbol{I} - \boldsymbol{A})^{-1}\,\boldsymbol{B} + D\right]\,U(s) \tag{1.122}$$

and the system transfer function is

$$G(s) = \boldsymbol{C}\,(s\,\boldsymbol{I} - \boldsymbol{A})^{-1}\,\boldsymbol{B} + D \tag{1.123}$$

This transfer function can be written in a different form using the determinant "det" and the adjoint matrix "adj":

$$G(s) = \boldsymbol{C}\,\frac{1}{\det(s\,\boldsymbol{I} - \boldsymbol{A})}\,[\mathrm{adj}(s\,\boldsymbol{I} - \boldsymbol{A})]\,\boldsymbol{B} + D \tag{1.124}$$

The determinant of $(s\,\boldsymbol{I} - \boldsymbol{A})$ is called the characteristic polynomial of $\boldsymbol{A}$ (dimension $n \times n$). Its n roots are the eigenvalues of $\boldsymbol{A}$.

1.10.2 Analog Representation

First Case

Consider a second-order system having the following transfer function

$$G_1(s) = \frac{K}{(\tau_1 s + 1)(\tau_2 s + 1)} \tag{1.125}$$

This transfer function is equivalent to the second-order ordinary differential equation

$$\tau_1\tau_2\frac{d^2y}{dt^2} + (\tau_1 + \tau_2)\frac{dy}{dt} + y = Ku(t) \tag{1.126}$$

which can be rewritten, in view of its further use in the analog block diagram, as

$$\frac{d^2y}{dt^2} = -a_0 y(t) - a_1\frac{dy}{dt} + b_0 u(t) \tag{1.127}$$

with:
$a_0 = 1/(\tau_1\tau_2)$, $a_1 = (\tau_1 + \tau_2)/(\tau_1\tau_2)$, $b_0 = K/(\tau_1\tau_2)$.

Note that the above differential equation is in terms of the deviation variables in the same way as for the transfer function. Let us assume that at

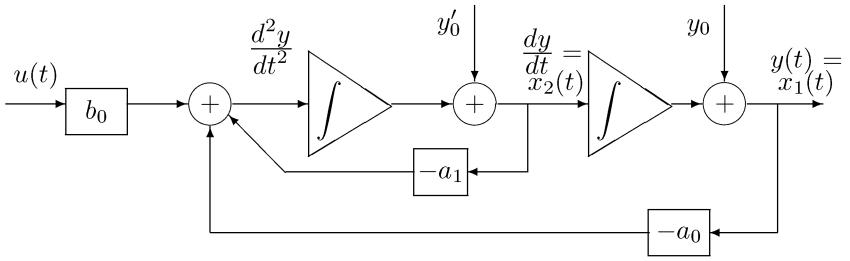


Figure 1.22: Analog block diagram of the state-space representation

time $t = 0$ the steady state prevails, $y(t = 0) = y(0)$ and $(dy/dt)_{t=0} = y'(0)$. The analog circuit representation of the above ordinary differential equation is shown in Fig. 1.22. The input $u(t)$ passes through a potentiometer (gain b_0), a summator, an integrator and another summator (supplying the initial condition $y'(0)$). The output from the second summator is passed through a second integrator and a third summator (supplying the initial condition $y(0)$). The final output is $y(t)$.
Designating $x_2(t)$ as the output of the first integrator and $x_1(t)$ as the output of the second integrator, the above system can be represented by the following state-space model

$$\begin{array}{lcl} \dot{x}_1(t) & = & x_2(t) \\ \dot{x}_2(t) & = & -a_0 x_1(t) - a_1 x_2(t) + b_0 u(t) \\ y(t) & = & x_1(t) \end{array} \tag{1.128}$$

or, using matrix form

$$\begin{array}{lcl} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} & = & \begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ b_0 \end{bmatrix} u(t) \\ y(t) & = & \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \end{array} \tag{1.129}$$

where $x_1(t)$ and $x_2(t)$ are the system state variables. The number of state variables is equal to the order of the differential equation or the order of the transfer function. The state of the system at any time t depends only on the initial conditions at t_0 and on the input $u(t)$ between t_0 and t. At initial time, the state variables are equal to the initial conditions and, in the following, they represent the evolution of the system.

Second Case

Consider a different second-order system having the following transfer function

$$G_2(s) = \frac{B(s)}{A(s)} = \frac{b_0\, s^2 + b_1\, s + b_2}{a_0\, s^2 + a_1\, s + a_2} \tag{1.130}$$

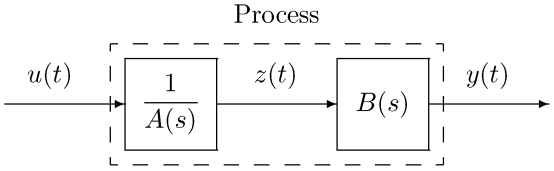


Figure 1.23: Representation of the partial state $z(t)$

This transfer function is not strictly proper. It is equivalent to the second-order ordinary differential equation

$$a_0 \frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_2\, \delta y(t) = b_0 \frac{d^2u}{dt^2} + b_1 \frac{du}{dt} + b_2\, \delta u(t) \tag{1.131}$$

The transformation into an analog form is not as straightforward as in the first case. We introduce the partial state $z(t)$ (Fig. 1.23) such that

$$Y(s) = B(s)\, Z(s) \qquad \text{and} \quad Z(s) = \frac{1}{A(s)}\, U(s) \tag{1.132}$$

As previously, we assume that $u(t)$, $z(t)$ and $y(t)$ are deviation variables. From the previous equations, we deduce

$$\begin{aligned} y(t) &= b_0 \frac{d^2z}{dt^2} + b_1 \frac{dz}{dt} + b_2\, z(t) \\ u(t) &= a_0 \frac{d^2z}{dt^2} + a_1 \frac{dz}{dt} + a_2\, z(t) \end{aligned} \tag{1.133}$$

Set

$$\begin{aligned} x_1(t) &= \frac{dz}{dt} \\ x_2(t) &= z(t) \end{aligned} \tag{1.134}$$

The state representation follows

$$\begin{aligned} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} &= \begin{bmatrix} \dfrac{-a_1}{a_0} & \dfrac{-a_2}{a_0} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} \dfrac{1}{a_0} \\ 0 \end{bmatrix} u(t) \\ y(t) &= \left[\left(\frac{-b_0\, a_1}{a_0} + b_1\right) \qquad \left(\frac{-b_0\, a_2}{a_0} + b_2\right) \right] \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \frac{b_0}{a_0}\, u(t) \end{aligned} \tag{1.135}$$

which allows to easily obtain the analog representation.

1.11 Dynamic Behaviour of Simple Processes

Initially, simple processes without a controller are considered (Fig. 1.24) and their open-loop behaviour is studied.
Let us consider the response of the system to two types of inputs $u(t)$:

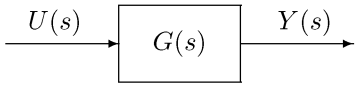


Figure 1.24: Open-loop block diagram of a process

- A unit step: $u = 1$ for $t > 0$, $u = 0$ for $t \leq 0$; the response of the system to this input is referred to as the step response.
- A Dirac unit impulse: $u = \delta$ (theoretical Dirac[1]); the response of the system to such an input is referred to as an impulse response.

Let $G(s)$ be the system transfer function subject to an input $U(s)$. Except for possible time delays, $G(s)$ is a rational fraction. For any physical input (impulse, step, ramp, sinusoidal, ...), $U(s)$ can also be expressed as a rational fraction, therefore

$$G(s) = \frac{N_g(s)}{D_g(s)} \quad ; \quad U(s) = \frac{N_u(s)}{D_u(s)} \tag{1.136}$$

The Laplace transform $Y(s)$ of the output can be decomposed into

$$Y(s) = G(s)\,U(s) = \frac{N_g(s)\,N_u(s)}{D_g(s)\,D_u(s)} = \frac{N_1(s)}{D_g(s)} + \frac{N_2(s)}{D_u(s)} = Y_n + Y_f \iff$$
$$\text{Response} = \text{Natural response} + \text{Forced response} \tag{1.137}$$

provided that the product $(G(s)\,U(s))$ is strictly proper and that denominators $D_g(s)$ and the $D_u(s)$ have no common roots.

The response $y_n(t)$ depends on the modes of $G(s)$ and is called the natural response of the system, while $y_f(t)$ depends on the modes of $U(s)$ (linked to the input type) and is referred to as the forced response of the system.

[1]The Dirac function $\delta(t)$ is defined by physicists as

$$\begin{aligned} &\delta(t) = 0 \qquad \forall t \neq 0 \\ &\delta(0) = +\infty \\ &\textstyle\int_{-\infty}^{+\infty} \delta(t) = 1 \end{aligned}$$

which is the limit of a real pulse function centred around 0 with unit area (strength or energy) and zero duration.

Mathematicians define the Dirac distribution such that the convolution of a function $f(x)$ by the Dirac distribution is equal to

$$f(x) * \delta(x) = f(x)$$

An important property of a Dirac distribution is

$$\int_{-\infty}^{+\infty} f(t)\delta(t - t_0) = f(t_0)$$

The Dirac distribution is equal to the derivative of a unit step function (Heaviside function).

1.11.1 First-Order Systems

A first-order system is described by a first-order differential equation of the form

$$\tau \frac{dy}{dt} + y(t) = K_p\, u(t) \tag{1.138}$$

The corresponding transfer function is equal to

$$G(s) = \frac{K_p}{\tau\, s + 1} \tag{1.139}$$

where τ is the time constant and K_p is the steady-state gain, or asymptotic gain of the process.

A first-order system can be represented by the block diagram shown in Fig. 1.25.

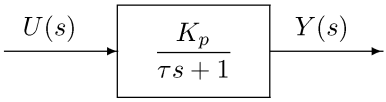


Figure 1.25: Block diagram of a first-order system

If the input u of the process is a step function with amplitude A, the Laplace transform of the input is

$$U(s) = \frac{A}{s} \tag{1.140}$$

Using the definition of the transfer function, the Laplace transform of the output is obtained

$$Y(s) = G(s)\, U(s) = \frac{K_p}{\tau\, s + 1}\, \frac{A}{s} = \frac{A\, K_p}{s} - \frac{A\, K_p\, \tau}{\tau\, s + 1} = Y_f(s) + Y_n(s) \tag{1.141}$$

and the time domain response (Fig. 1.26) is

$$y(t) = A\, K_p\, (1 - \exp(-t/\tau)) \tag{1.142}$$

The forced and natural parts of the response are respectively equal to

$$y_f(t) = A\, K_p \qquad ; \quad y_n(t) = -A\, K_p\, \exp(-t/\tau)) \tag{1.143}$$

With respect to the input of amplitude A, the asymptotic output (when $t \to \infty$) is thus multiplied by the gain of the process K_p. A first-order process is also called a "first-order lag".

The time constant τ corresponds to the time necessary for the system response to reach 63.2% of its asymptotic value for a step input. After 2τ, the response reaches 86.5% and after 5τ it reaches 99.3% (Table 1.2).

Several real physical systems have first-order dynamics. Examples of such systems are:

- Systems storing mass, energy or momentum,
- Systems showing resistance to the flow of mass, energy or momentum.

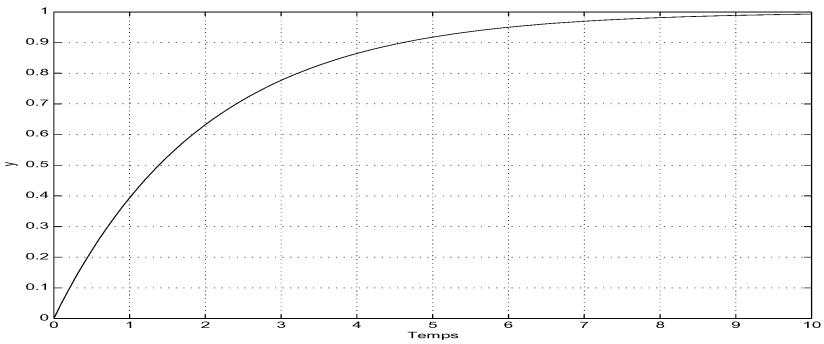


Figure 1.26: Response of a first-order system ($K_p = 1, \tau = 2$) to a unit step function

Table 1.2: Response of a first-order system to a unit step function expressed in percentage of the asymptotic value

Time	Percentage of the asymptotic value
0	0
τ	63.21%
2τ	86.47%
3τ	95.02%
4τ	98.17%
5τ	99.33%
6τ	99.75%
7τ	99.91%

1.11.2 Integrating Systems

Integrating or pure capacitive processes are those whose dynamics only contain the first-order derivative of $y(t)$

$$\frac{dy}{dt} = K_p\, u(t) \tag{1.144}$$

The corresponding transfer function is

$$G(s) = \frac{K_p}{s} \tag{1.145}$$

The Laplace transform of the output of such a system to a step function with magnitude A is

$$Y(s) = \frac{A\, K_p}{s^2} \tag{1.146}$$

The time domain response $y(t)$ (Fig. 1.27) is thus equal to

$$y(t) = A\, K_p\, t \tag{1.147}$$

The process is referred to as a "pure capacitive" or a "pure integrator". The term "capacitive" signifies the accumulation of electrical charges, energy or mass. A surge tank can behave as a pure capacitive process.

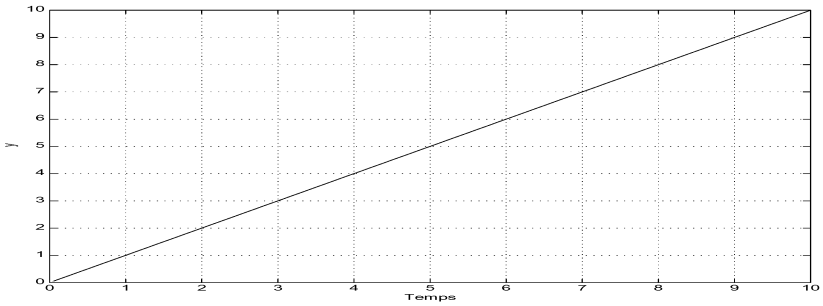


Figure 1.27: Response of a pure capacitive system ($K_p = 1$) to a unit step input

1.11.3 Second-Order Systems

A second-order system is described by a second-order differential equation written in the classical form as

$$\tau^2 \frac{d^2y(t)}{dt^2} + 2\,\zeta\,\tau \frac{dy(t)}{dt} + y(t) = K_p\,u(t) \tag{1.148}$$

with the corresponding transfer function

$$G(s) = \frac{K_p}{\tau^2\,s^2 + 2\,\zeta\,\tau\,s + 1} \tag{1.149}$$

where τ is the natural period of oscillation of the system which determines the stabilization time of the system, ζ is the damping coefficient and K_p is the steady-state gain of the system.
The notions of natural period of oscillation and of damping factor are related to the damped or undamped oscillators. For $\zeta = 0$, the expression (1.156) shows that the response to a step input oscillates continuously with a frequency $1/\tau$ in radians/time unit.

The transfer function of a second-order system is sometimes written as

$$G(s) = \frac{K_p\,\omega_n^2}{s^2 + 2\,\zeta\,\omega_n\,s + \omega_n^2} \tag{1.150}$$

where $\omega_n = 1/\tau$ is the natural undamped frequency and $\sigma = \zeta\,\omega_n$ is the damping parameter.

Several real physical processes exhibit second-order dynamics, among them are:

– Two first-order systems in series.

– Intrinsic second-order systems, e.g. mechanical systems having an acceleration.
– Feedback or closed-loop transfer function of a first-order process with a PI controller.

Note that the transfer function $G(s)$ defined by Eq. (1.149) has two poles, roots of: $\tau^2\, s^2 + 2\,\zeta\,\tau\, s + 1 = 0$, which are equal to

$$\begin{cases} s_i = \dfrac{1}{\tau}\,(-\zeta \pm \sqrt{\zeta^2 - 1}) & \text{if: } \zeta \geq 1 \\ \qquad\; \dfrac{1}{\tau}\,(-\zeta \pm j\,\sqrt{1-\zeta^2}) = \omega_n(-\zeta \pm j\sqrt{1-\zeta^2}) = -\sigma \pm j\,\omega_a & \text{if: } 0 \leq \zeta \leq 1 \end{cases} \tag{1.151}$$

If the natural period of oscillation τ is fixed, then the position of the poles depends only on the damping coefficient ζ. The shape of the open-loop response to a given input is determined by the location of these poles on the s-plane. For $0 \leq \zeta \leq 1$, the natural frequency ω_n is equal to the distance of the poles from the origin, the damped frequency ω_a is equal to the distance of the poles from the real axis, and the damping parameter σ is equal to the distance of the poles from the imaginary axis.

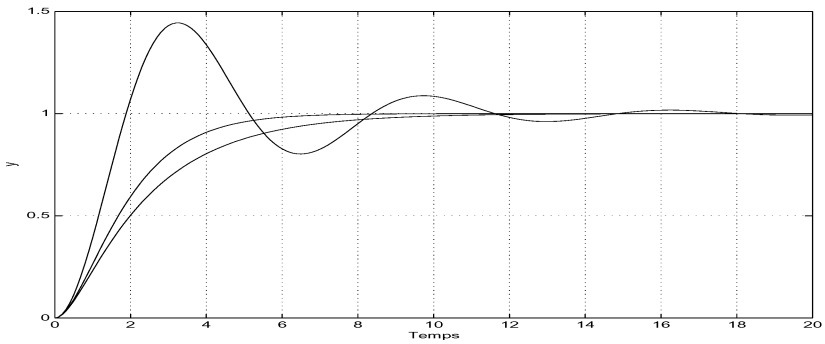


Figure 1.28: Normalized response of a second-order system to a unit step function for different values of the damping coefficient ζ (= 0.25 ; 1 ; 1.3 resulting in oscillatory underdamped response to overdamped response) ($K_p = 1, \tau = 1$)

If the input is a step function with magnitude A, the output Laplace transform is equal to

$$Y(s) = A\,\frac{K_p}{s\;(\tau^2\, s^2 + 2\,\zeta\,\tau\, s + 1)} \tag{1.152}$$

which can be decomposed into

$$Y(s) = \frac{A\,K_p}{s} - \frac{A\,K_p\,\tau^2\, s + 2\,\zeta\,\tau}{\tau^2\, s^2 + 2\,\zeta\,\tau\, s + 1} = Y_f(s) + Y_n(s) \tag{1.153}$$

The overall response consists of the forced and the natural responses

$$y(t) = y_f(t) + y_n(t) \tag{1.154}$$

The forced response is equal to

$$y_f(t) = A\,K_p \tag{1.155}$$

and the overall response is

$$y(t) = \begin{cases} AK_p \left\{1 - \exp(-\zeta t/\tau)\ \left[\cos\left(\frac{\sqrt{1-\zeta^2}}{\tau}t\right) + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\left(\frac{\sqrt{1-\zeta^2}}{\tau}t\right)\right]\right\} \\ \text{if: } 0 \le \zeta < 1 \\ AK_p \left[1 - (1 + \frac{t}{\tau})\exp(-t/\tau)\right] \\ \text{if: } \zeta = 1 \\ AK_p \left\{1 - \exp(-\zeta t/\tau)\ \left[\cosh\left(\frac{\sqrt{\zeta^2-1}}{\tau}t\right) + \frac{\zeta}{\sqrt{\zeta^2-1}}\sinh\left(\frac{\sqrt{\zeta^2-1}}{\tau}t\right)\right]\right\} \\ \text{if: } 1 < \zeta \end{cases} \tag{1.156}$$

The forced response is constant and equal to AK_p while the natural response tends towards 0 when $t \to \infty$. The natural response takes into account the natural modes of the system and thus depends on the value of ζ (Fig. 1.28):

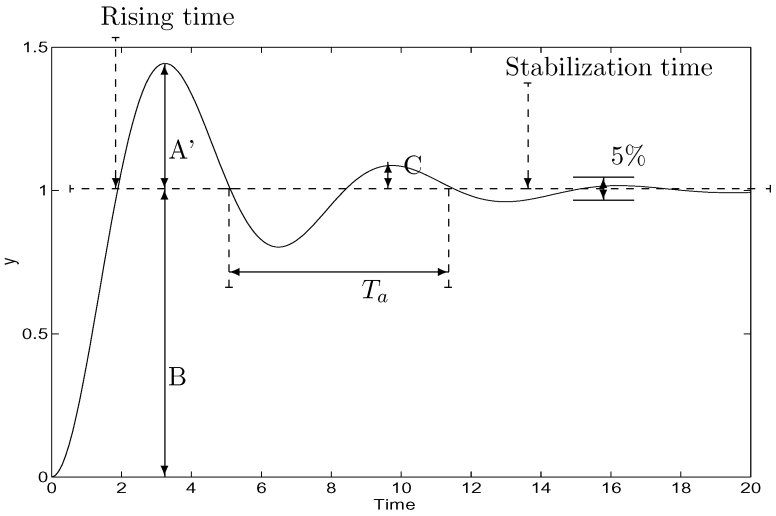


Figure 1.29: Response of a second-order system to a unit step input

- For $\zeta > 1$, there will be two real and distinct poles. The response is overdamped (multi-capacitive systems) with no overshoot.
- For $\zeta = 1$, there will be one multiple second-order pole. The response is critically damped, which corresponds to the faster overdamped response.
- For $0 < \zeta < 1$, there will be two complex conjugate poles with negative real part. The response is underdamped. This response is initially faster than the critically damped and overdamped responses, which are sluggish; the drawback is the resulting overshoot.

With reference to the underdamped response of Fig. 1.29, the following terms are defined:

- Overshoot

$$\text{overshoot} = \frac{A'}{B} = \exp\left(\frac{-\pi\,\zeta}{\sqrt{1-\zeta^2}}\right) \tag{1.157}$$

- Decay ratio $= C/A' = (\text{overshoot})^2$
- Natural period of oscillation is defined for a system with a damping coefficient ζ equal to zero. Such a system oscillates continuously with the natural period $T_n = 2\,\pi\,\tau = 2\,\pi/\omega_n$ and the undamped natural frequency ω_n.
- Actual period of oscillation T_a, which is the time between two successive peaks, characterized by its damped frequency ω_a

$$\omega_a = \omega_n\,\sqrt{1-\zeta^2} = \frac{\sqrt{1-\zeta^2}}{\tau} = \frac{2\,\pi}{T_a} \tag{1.158}$$

- Rise time: this is the time necessary to reach the asymptotic value for the first time

$$t_m = \frac{1}{\omega_n\sqrt{1-\zeta^2}}\,\text{arctg}\left(-\frac{\sqrt{1-\zeta^2}}{\zeta}\right) \tag{1.159}$$

  It can also be defined as the time necessary to go from 10% to 90% of the asymptotic value and, in that case, it can be approximated (Goodwin and Sin, 1984) by

$$t_m \approx \frac{2.5}{\omega_n} \tag{1.160}$$

- First peak reach time: the time necessary for the response to reach the first peak

$$t_p = \frac{\pi}{\omega_d} = \frac{\pi}{\omega_n\sqrt{1-\zeta^2}} \tag{1.161}$$

- Settling time: time necessary for the response to remain in an interval between $\pm\epsilon$ ($\pm 5\%$, or $\pm 2\%$) of the asymptotic value. For $\pm\epsilon = \pm 1\%$, according to Goodwin and Sin (1984), the settling time is

$$t_s \approx \frac{4.6}{\zeta\omega_n} \tag{1.162}$$

For $(0 < \zeta < 1)$, the time domain response can be written as

$$y(t) = AK_p - AK_p\,\frac{\omega_n}{\omega_a}\,\exp(-\sigma\,t)\,\sin(\omega_d\,t + \theta) \tag{1.163}$$

with

$$\theta = \arccos\zeta = \arctan\left(\sqrt{\frac{1-\zeta^2}{\zeta}}\right) = \arcsin(\sqrt{1-\zeta^2}) \tag{1.164}$$

and the envelope of the undamped sinusoidal response is

$$\exp(-\sigma\,t)\,\sin(\omega_d\,t + \theta) \tag{1.165}$$

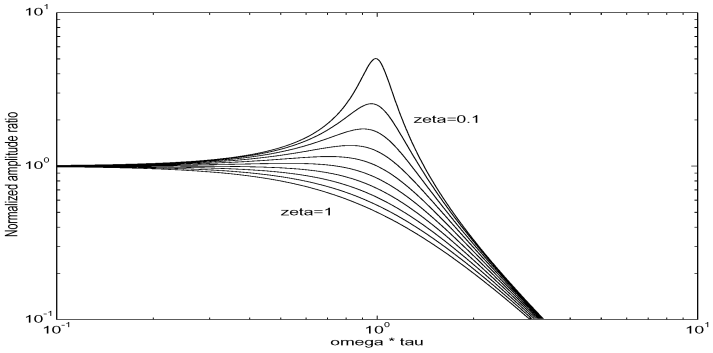


Figure 1.30: Normalized amplitude ratio for a sinusoidal input with varying damping coefficient ζ between 0 and 1 per increment of 0.1

The time domain response of a second-order system subjected to a sinusoidal input: $u(t) = A\ \sin(\omega t)$, after the transient response decays, will take the form

$$y_\infty(t) = \frac{K_p\ A}{\sqrt{[1-(\omega\tau)^2]^2 + (2\zeta\omega\tau)^2}} \sin(\omega t - \text{arctang}\left[\frac{2\zeta\omega\tau}{1-(\omega\tau)^2}\right]) \qquad (1.166)$$

and the normalized amplitude ratio is equal to

$$RA_n = \frac{1}{\sqrt{[1-(\omega\tau)^2]^2 + (2\zeta\omega\tau)^2}} \qquad (1.167)$$

which is maximum at a frequency $\omega_{\max}$ given by

$$\omega_{\max} = \sqrt{(1-2\zeta^2)}/\tau \qquad (1.168)$$

The normalized amplitude ratio has a maximum equal to $1/(2\zeta\sqrt{1-\zeta^2})$ for $0 \leq \zeta \leq 0.707$. This maximum increases very quickly when ζ becomes small (Fig. 1.30).

Large oscillations are not desired, therefore small damping coefficients ζ must be avoided. In controlled processes, a damping coefficient around $\zeta = 1$ (conservative) or $\zeta = 0.7$ (low overshoot, fast response), is often recommended.

1.11.4 Higher-Order Systems

Three types of higher-order systems will be described:

- n First-order processes in series (multi-capacitive).
- Processes with time delay.
- Processes with inverse response.

n First-Order Processes in Series

The transfer function of n first-order processes in series is obtained by multiplying the transfer functions of n first-order systems

$$G(s) = \prod_{i=1}^{n} \frac{K_{pi}}{\tau_i s + 1} \tag{1.169}$$

Example 1.3: **Tubular Plug Flow Reactor**

Consider a tubular plug flow reactor with a mean residence time equal to τ_1. In the absence of reaction, for a simple flow, the outlet concentration is equal to the inlet one, just delayed by the residence time τ_1. Suppose now that a first-order reaction A $\rightarrow$ B with a reaction rate $r_A = k\,C_A$ is carried out in the reactor, the reactor being fed with a reactant stream at inlet concentration C_{A_0}. This represents a distributed-parameter system whose model is given by a partial differential equation. Another approach to model the reactor is to discretize it into n elementary reactors (Fig. 1.31) with a residence time given by

$$\tau_n = \frac{\tau_1}{n} \tag{1.170}$$

The component mass balance on each element is given by

$$\frac{d\,C_A(i)}{dt} = \frac{1}{\tau_n}\left[C_A(i-1) - C_A(i)\right] - k\,C_A(i) \tag{1.171}$$

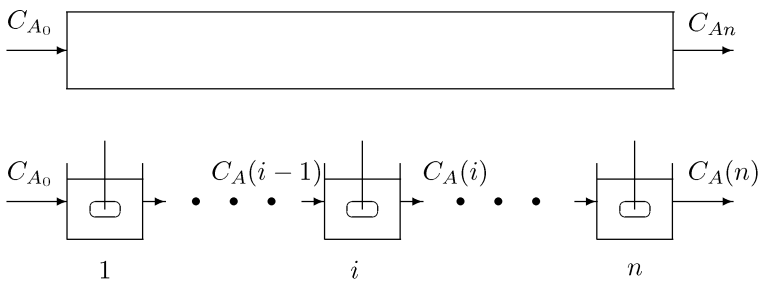


Figure 1.31: Decomposition of a tubular reactor into a series of n continuous perfectly stirred tank reactors

In the absence of reaction, the flow is simply represented by n first-order systems in series with unit gain

$$\frac{\bar{C}_A(i)}{\bar{C}_A(i-1)} = \frac{1}{\tau_n\, s + 1} \tag{1.172}$$

The observed response at the reactor exit is similar to an overdamped system with a sluggish sigmoidal shape. Figure 1.32 represents the response of the system to a step increase in C_{A_0} from 0.4 to 0.5, using different numbers of

discretization elements n ranging from 1 to 100. Note that as n approaches 100, the response looks like the input step with a delay time equal to the mean residence time of the tubular plug flow reactor (assumed 6 min. in this simulation).

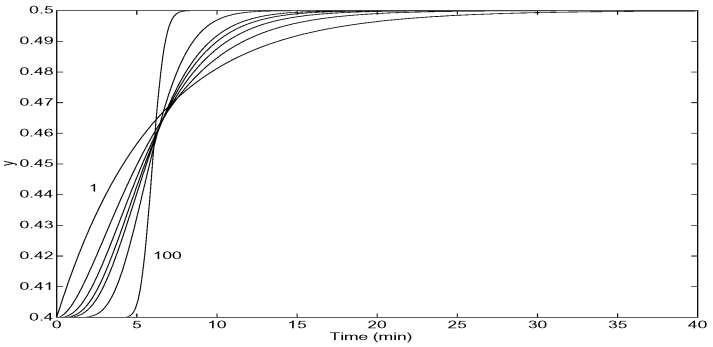


Figure 1.32: Response of the discretized plug flow reactor to a step increase in the inlet concentration from 0.4 to 0.5 with the number n of discretized elements, ranging from $1, 2, 3, 4, 5, 10$ and 100 (in the absence of reaction)

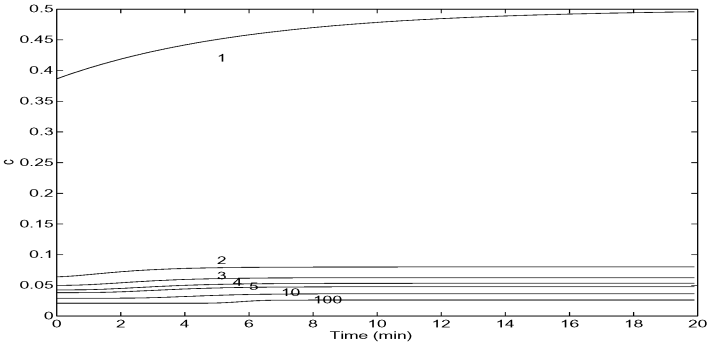


Figure 1.33: Response of the discretized plug flow reactor to a step increase in the inlet reactant concentration from 0.4 to 0.5 in the presence of a first-order reaction

In the case where a chemical reaction occurs, each element represents a CSTR. Except in the simple case where the reaction is first-order, the model is in general nonlinear and would need a linearization around a steady state. For a first-order reaction, the transfer function is

$$\frac{\bar{C}_A(i)}{\bar{C}_A(i-1)} = \frac{1}{\tau_n\, s + 1 + k\, \tau_n} \tag{1.173}$$

Note that the process gain is no longer unity due to the depletion of reactant by chemical reaction. The process response to a unit increase in the inlet concen-

tration of reactant C_{A_0} from 0.4 to 0.5 in the presence of reaction is shown in Fig. 1.33. Due to the chemical reaction, the asymptotic outlet concentrations decrease with the number of elements of discretization in a similar manner to a tubular reactor. In comparison with Fig. 1.32, the response is more sluggish.

Another example of a series of elementary systems resulting in an overall higher order system is the staged processes such as a tray distillation column. Each tray of a distillation column can be often considered as a first-order interacting process.

A packed column (distillation, absorption or chromatography) can be modelled by partial differential equations and could be discretized in a similar manner to a tubular reactor.

The dynamics of distributed-parameter systems and higher-order systems are often approximated by a first- or second-order overdamped model with time delay. For example, in a distillation column, a set of plates such as the stripping or the enrichment section is often identified as first-order with delay.

Processes with Time Delay

Time delay may be an inherent dynamic characteristic of a process or due to the measurement. In the former case, the process input does not immediately affect the process output. In the latter case, the measured signal received by the controller does not correspond to the contemporary process information and suffers from delay. A common example of time delay is the transportation lag, which may be either due to the process or measurement or both. Consider, for example, the concentration measurement in a reactor which frequently is not done in situ (Fig. 1.34). The measuring device is mounted on a sampling loop. The sample is pumped through the loop and experiences some transportation lag t_d to reach the sensor.

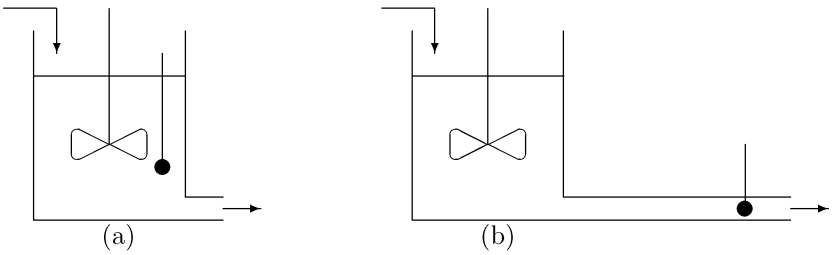


Figure 1.34: Case (a): In situ sensor. Case (b): sensor placed in the exit pipe, inducing a transportation lag

In the case of a distillation column, in general distillate and bottom concentrations are controlled by manipulating, for example, the reflux flow and the steam

flow to the reboiler. Measurements are typically the levels in the reboiler and in the condenser, temperatures at different points of the column, and distillate and bottom concentrations. Temperature and level measurements can be considered as instantaneous. This is not the case for concentration measurement, e.g. in the case of refineries, a chromatograph some distance away from the distillation column is often used. In such cases, the delay time will consist of a transportation lag to pump the sample from the process to the analyzer and an additional time for the analysis of the sample, which in the case of a chromatograph could be in the order of several tens of seconds. A sample is taken from the distillation column at time t_1 and the result is available at time t_2, where $t_2 - t_1 = t_d$. During the measurement, the process continues to evolve.

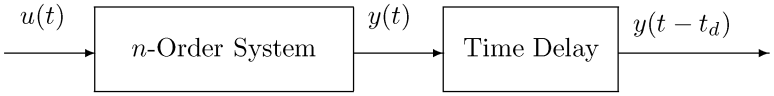


Figure 1.35: Representation of a system with time delay

The time delay poses a problem in process control. The time delay between an input and an output (Fig. 1.35) means that the input variations have no immediate influence on the output.

For a first-order system with a time delay, the transfer function linking the input $u(t)$ and the delayed output $y(t - t_d)$ is

$$\frac{\mathcal{L}[y(t - t_d)]}{\mathcal{L}[u(t)]} = \frac{K_p \exp(-t_d\, s)}{\tau\, s + 1} \tag{1.174}$$

The exponential term is a nonlinear term. It is often approximated, for example, by a Padé approximation (here a first-order approximation), which converts the delay term to a rational fraction

$$\exp(-t_d\, s) \approx \frac{1 - \dfrac{t_d}{2}\, s}{1 + \dfrac{t_d}{2}\, s} \tag{1.175}$$

A more accurate approximation of the time delay is realized by the second-order Padé approximation.

$$\exp(-t_d\, s) \approx \frac{1 - \dfrac{t_d}{2}\, s + \dfrac{t_d^2}{12}\, s^2}{1 + \dfrac{t_d}{2}\, s + \dfrac{t_d^2}{12}\, s^2} \tag{1.176}$$

Figure 1.36 shows the unit step response of a first-order system without time delay $y(t)$, with exact time delay $y(t - t_d)$, with first-order Padé approximation y_1 and with second-order Padé approximation y_2. Note that the approximations are valid for times much larger than the time delay. Initially both first- and second-order approximations exhibit inverse response due to the zeros of the

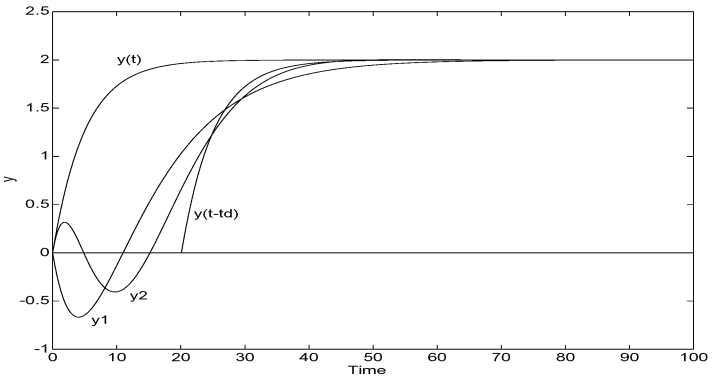


Figure 1.36: Response of a first-order system ($K_p = 2, \tau = 5$) with a time delay of 20 s to a step input: $y(t)$ (without time delay), $y(t - t_d)$ (with time delay), y_1 (Padé first-order), y_2 (Padé second-order)

rational transfer functions introduced by Padé approximations. The number of intersections of the approximated response with the time axis is equal to the order of Padé approximation and corresponds to the number of positive real zeros or complex zeros with positive real part of the transfer function, i.e. for the first-order Padé approximation, there is one real positive zero and for the second-order Padé approximation, there are two conjugate complex zeros with positive real part.

There exists no perfect approximation for the time delay. However, digital computers handle time delays with relative ease, in particular in the case of digital control.

Processes with Inverse Response

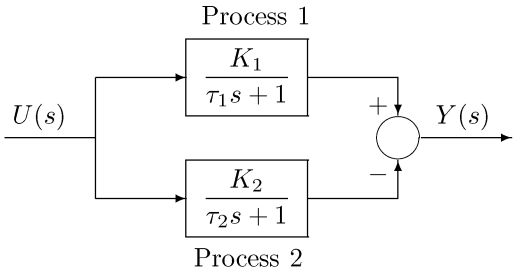


Figure 1.37: Representation of a system with inverse response

In a system with inverse response such as that represented in Fig. 1.37, the overall response y_g (Fig. 1.38) initially moves in a direction opposite to its

final direction. This behaviour is caused by the addition of two opposing subprocesses. At the beginning, process 2, which has a smaller time constant, is dominant. Subsequently, process 1, which has a higher gain ($K_1 > K_2$), becomes dominant. The overall process is thus composed of two competitive processes having different time constants and different gains.

To produce an inverse response, it is necessary and sufficient that the plant transfer function exhibits a positive real zero or a complex zero with a positive real part. For the example shown in Fig. 1.37, the overall transfer function will have a positive zero if

$$\frac{K_2}{\tau_2} > \frac{K_1}{\tau_1} \tag{1.177}$$

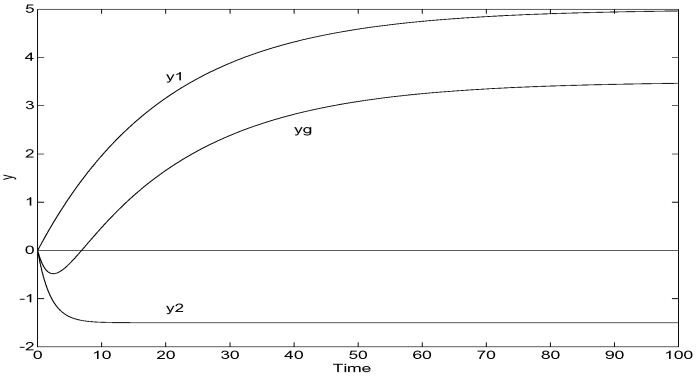


Figure 1.38: Inverse response of the system shown in Fig. 1.37 to a unit step input ($K_1 = 5, \tau_1 = 20, K_2 = 1.5, \tau_2 = 2$)

Chemical processes which exhibit an inverse response are not rare. For example, an increase in the vapour boil-off (increase in the steam flow rate to the reboiler) in a distillation column results in an inverse response in the liquid level in the bottom of the column. The increase in the vapour boil-off, initially, decreases the liquid level in the bottom of the column. Subsequently, due to the increased vapour flow rate and frothing on the trays immediately above the reboiler, the liquid level may eventually increase.

Another example is the response in the liquid level in a boiler due to an increase in the inlet water flow rate. The initial increase in the water flow rate decreases the liquid level because of the collapse of the vapour bubbles. However, if the steam production is maintained at a constant rate, the liquid level will eventually increase.

A third example is the response in the exit temperature of a fixed-bed reactor with exothermic reactions to a step increase in the feed temperature. If the feed temperature increases, the rate of reactions at the reactor inlet will increase and the reactants are consumed close to the reactor inlet. This will move the hot spot closer to the reactor entrance and consequently the exit re-

actor temperature decreases. However, eventually the reactor exit temperature will increase as a result of the increase in the inlet temperature.

1.11.5 Process Identification in the Continuous Domain

General Principles

It is possible to develop a model for an existing process using only the input-output data. This empirical technique for developing process models is referred to as process or system identification (Sinha and Rao, 1991; Unbehauen and Rao, 1987; Walter and Pronzato, 1997; Young, 1981). System identification for complex processes requires less engineering effort compared to the theoretical model development. Of course, the application field of such an identified model is more limited. In open-loop system identification, the controller is switched to manual, the controller output is changed by a step function or a series of pseudo-random-binary-signal (PRBS) or any other exciting input sequence, and the process response is monitored. The input-output data are used to develop the system model. This model will represent the dynamics of the combination of the final control element (the control valve), the process, and the measuring element. Measurement noise will generally be superposed on the actual process response. Sophisticated process identification techniques are capable of distinguishing the process model from the noise model.

Often, a low-order linear model based on the a priori information of the process is assumed, i.e. the structure of the model is fixed. Under these conditions, process identification is reduced to the determination of the unknown model parameters, i.e. a parameter estimation problem. Let us assume that the process model is given by

$$y(t) = f(t, \boldsymbol{\theta}) \tag{1.178}$$

where $\boldsymbol{\theta}$ is the unknown parameter vector. In order to determine the unknown parameter vector, one may use the least-squares technique by minimizing the following objective function, which is the sum of the squares of errors between the measured output y_i at time t_i and the output predicted by the model $\hat{y}_i$

$$J(\boldsymbol{\theta}) = \sum_i (y_i - \hat{y}_i)^2 \tag{1.179}$$

Other criteria can be introduced in particular by the use of weighting factors.

Two main cases are distinguished depending on whether the model is linear or nonlinear. In the case of a linear model with respect to m parameters θ_j, the model can be written as

$$y(\boldsymbol{x}) = \sum_{j=1}^{m} \theta_j \, \phi_j(\boldsymbol{x}) \tag{1.180}$$

A set of n input-output observations can be collected in a matrix Φ such that

$$\Phi_{ij} = \phi_j(\boldsymbol{x}_i) \quad ; \quad 1 \leq i \leq n \quad ; \quad 1 \leq j \leq m \tag{1.181}$$

The parameter vector estimated by minimization of the criterion (1.179) is equal to

$$\hat{\boldsymbol{\theta}} = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{y} \tag{1.182}$$

where $\boldsymbol{y}$ is the measured vector.

Example 1.4: Identification of a Linear Model

Consider the example of determining the parameters in the model of the heat capacity of a fluid expressed with respect to temperature by

$$C_p = a + b\,T + c\,T^2 \tag{1.183}$$

This model is nonlinear with respect to temperature but linear with respect to the coefficients a, b, c. As we wish to determine the coefficients, we must realize at least three experiments. In the case where only three experiments are performed, the coefficients are the solution of a perfectly determined system which can be solved by the Gauss method, for example. In the case where more than three experiments are performed, it becomes a least-squares problem. Assume that we perform $n > 3$ experiments which give heat capacities $C_{p,i}$ at temperatures T_i. The matrix Φ and vector $\boldsymbol{y}$ of Eq. (1.182) are equal to

$$\Phi = \begin{bmatrix} 1 & T_1 & T_1^2 \\ 1 & T_2 & T_2^2 \\ \vdots & \vdots & \vdots \\ 1 & T_n & T_n^2 \end{bmatrix} \quad ; \quad \boldsymbol{y} = \begin{bmatrix} C_{p,1} \\ C_{p,2} \\ \vdots \\ C_{p,n} \end{bmatrix} \tag{1.184}$$

The vector of estimated parameters is then

$$\begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{bmatrix} = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{y} \tag{1.185}$$

In the case of a nonlinear model, which is most common in practice, more general optimization methods such as direct search or gradient-type methods such as the generalized reduced gradient method or the quasi-Newton method (Fletcher, 1991; Gill et al., 1981), must be employed. Powerful algorithms exist for solving such problems as the BFGS method (Byrd et al., 1995; Zhu et al., 1994) or sequential quadratic programming (SQP) under NLPQL form (Schittkowski, 1985). It is also possible to linearize the non-linear model with respect to the parameter vector and then use a linear parameter estimation method.

Example 1.5: Identification of a Nonlinear Model

Consider the example of determining the parameters in the model of the saturated vapour pressure of a fluid expressed with respect to temperature by Antoine's law

$$\ln(P_{sat}) = A - \frac{B}{C + T} \tag{1.186}$$

We can make remarks similar to the previous case. This model is nonlinear with respect to the temperature and nonlinear with respect to the coefficients A, B, C. In the case where only three experiments are performed, the coefficients are the solution of a perfectly determined system which is now nonlinear and could be solved by the Newton-Raphson method (Carnahan et al., 1969). In the case where more than three experiments are performed, it becomes a least-squares problem. Assuming that we make n experiments which give vapour pressures P_i at temperatures T_i, a least-squares criterion to be minimized is written as

$$J = \sum_{i=1}^{n} (\ln(P_i) - A + \frac{B}{C + T_i})^2 \tag{1.187}$$

whose minimization with respect to parameters can be performed by a quasi-Newton method.

A First-Order Model

If a time delay is present, first it must be estimated and then the system is examined without this delay. The steady-state gain K_p is also estimated from the asymptotic response. The time constant τ can be evaluated by using Table 1.2, for example, by searching the time at which 63.2% of the asymptotic value is reached. For a step change with magnitude A in the input, the time domain response is given by

$$\ln(1 - \frac{y}{K_p\, A}) = -\frac{t}{\tau} \tag{1.188}$$

which corresponds to a straight line with a slope $-1/\tau$.

A Second-Order Model

Similar to a first-order model, the time delay must be estimated first, then the steady-state gain K_p.

If the second-order system is overdamped, its transfer function can be written as

$$G(s) = \frac{K_p}{(\tau_1\, s + 1)\,(\tau_2\, s + 1)} \tag{1.189}$$

with two unknown time constants. Harriott's method offers a simple graphical method for the determination of the unknown time constants based on the measurements of process outputs at two points during its evolution. A more accurate method is the numerical non-linear optimization using the entire process output response to a step change in the input

$$y(t) = A\, K_p \left[1 + \frac{\tau_1}{\tau_2 - \tau_1} \exp(-t/\tau_1) + \frac{\tau_2}{\tau_1 - \tau_2} \exp(-t/\tau_2)\right] \tag{1.190}$$

When plotted versus time, this function has a sigmoidal shape with an inflection point at

$$t = \frac{\tau_1 \tau_2}{\tau_1 - \tau_2} \ln(\frac{\tau_1}{\tau_2}) \tag{1.191}$$

If the second-order system is underdamped, its transfer function is

$$G(s) = \frac{K_p}{\tau^2 \, s^2 + 2 \, \zeta \, \tau s + 1} \tag{1.192}$$

Two parameters must be estimated, the damping coefficient and the time constant τ. Graphical methods can be used to determine the decay ratio or the overshoot to estimate the unknown parameters.

However, a better approach is the nonlinear optimization using the response given by Eq. (1.156) to minimize a criterion such as the one given in Eq. (1.179). Consider the response of the model denoted by $y_{mod,i}$ and the experimental response denoted by $y_{exp,i}$ at n different instants t_i. The nonlinear optimization problem is expressed as

$$\min_{\tau, \zeta, K_p} \sum_{i=1}^{n} (y_{exp,i} - y_{mod,i})^2 \tag{1.193}$$

which again can be solved by a quasi-Newton-type method. The response of the model depends analytically on the parameters and the criterion can be analytically differentiated. If this differentiation seems too difficult, it may be performed numerically.

The Method of Moments

The advantage of the method of moments is that it can be used with any type of input.

This method is based on the definition of the Laplace transform of the impulse response $g(t)$ of a system, which is its transfer function

$$G(s) = \int_0^\infty \exp(-s \, t) \, g(t) \, dt \tag{1.194}$$

As the n-th-order moment of a function $f(x)$ is defined by

$$\mathcal{M}_n(f) = \int_0^\infty x^n \, f(x) \, dx \tag{1.195}$$

it can be noticed that the first two derivatives of $G(s)$ with respect to s

$$G'(s) = -\int_0^\infty t \, \exp(-s \, t) \, g(t) \, dt \quad ; \quad G''(s) = \int_0^\infty t^2 \, \exp(-s \, t) \, g(t) \, dt \tag{1.196}$$

are related to the moments of the impulse response function by

$$G(0) = \int_0^\infty g(t) \, dt \quad ; \quad G'(0) = -\int_0^\infty t \, g(t) \, dt \quad ; \quad G''(0) = \int_0^\infty t^2 \, g(t) \, dt \tag{1.197}$$

Thus $G(0)$, $-G'(0)$, $G''(0)$ are respectively the zero-, first- and second-order moments of the impulse response $g(t)$.
Note that the above three integrals can be calculated by using the measured output response. Consider the following systems for the application of this method:
• A first-order model with time delay

$$G(s) = \frac{K_p \, \exp(-t_d \, s)}{\tau \, s + 1} \tag{1.198}$$

giving

$$G(0) = K_p \quad ; \quad G'(0) = -K_p \, (\tau + t_d) \quad ; \quad G''(0) = K_p \, (2 \, \tau^2 + 2 \, \tau \, t_d + t_d^2) \tag{1.199}$$

from which the three unknown parameters K_p, τ and t_d can be determined.
• An overdamped second-order model with time delay (previously determined)

$$G(s) = \frac{K_p \, \exp(-t_d \, s)}{(\tau_1 \, s + 1) \, (\tau_2 \, s + 1)} \tag{1.200}$$

from which

$$\begin{array}{ll} G(0) & = K_p \quad ; \quad G'(0) = -K_p \, (\tau_1 + \tau_2 + t_d) \\ G''(0) & = K_p \, \left[(\tau_1 + \tau_2 + t_d)^2 + \tau_1^2 + \tau_2^2 \right] \end{array} \tag{1.201}$$

The time delay can be obtained by inspection of the response curve and the other parameters are obtained from the moments.
• An underdamped second-order model with time delay (previously determined)

$$G(s) = \frac{K_p \, \exp(-t_d \, s)}{\tau^2 \, s^2 + 2 \, \zeta \, \tau \, s + 1} \tag{1.202}$$

from which

$$\begin{array}{ll} G(0) & = K_p \quad ; \quad G'(0) = -K_p \, (2 \, \zeta \, \tau + t_d) \\ G''(0) & = K_p \, \left[t_d^2 + 4 \, \zeta \, \tau \, t_d - 2 \, \tau^2 + 8 \, \zeta^2 \, \tau^2 \right] \end{array} \tag{1.203}$$

Note that in this method, it is possible to use any type of input. We have, in general

$$\begin{array}{ll} Y(s) & = G(s) \, U(s) \quad ; \quad Y'(s) = G'(s) \, U(s) + G(s) \, U'(s) \\ Y''(s) & = G''(s) \, U(s) + G(s) \, U''(s) + 2 \, G'(s) \, U'(s) \end{array} \tag{1.204}$$

from which the following equations are deduced

$$\begin{array}{ll} Y(0) & = G(0) \, U(0) \quad ; \quad Y'(0) = G'(0) \, U(0) + G(0) \, U'(0) \\ Y''(0) & = G''(0) \, U(0) + G(0) \, U''(0) + 2 \, G'(0) \, U'(0) \end{array} \tag{1.205}$$

These quantities can be calculated by the following equations:

$$U(0) = \int_0^\infty u(t) \, dt \quad ; \quad U'(0) = -\int_0^\infty t \, u(t) \, dt \quad ; \quad U''(0) = \int_0^\infty t^2 \, u(t) \, dt \tag{1.206}$$

and

$$Y(0) = \int_0^\infty y(t)\,dt \quad ; \quad Y'(0) = -\int_0^\infty t\,y(t)\,dt \quad ; \quad Y''(0) = \int_0^\infty t^2\,y(t)\,dt \tag{1.207}$$

In order to determine the characteristic parameters of the system, the moments method uses the entire input and output curves and can be applied to any type of the input signal. This method should be preferred to any method based on only two given points of an output response.

References

R.B. Bird and E.N. Lightfoot. *Transport Phenomena*. Wiley, New York, 1960.

P. Borne, G. Dauphin-Tanguy, J.P. Richard, F. Rotella, and I. Zambettakis. *Commande et Optimisation des Processus*. Technip, Paris, 1990.

P. Borne, G. Dauphin-Tanguy, J.P. Richard, F. Rotella, and I. Zambetakis. *Commande et Optimisation des Processus*. Technip, Paris, 1992a.

P. Borne, G. Dauphin-Tanguy, J.P. Richard, F. Rotella, and I. Zambetakis. *Modélisation et Identification des Processus*, volume 1, 2. Technip, Paris, 1992b.

P. Borne, G. Dauphin-Tanguy, J.P. Richard, F. Rotella, and I. Zambetakis. *Analyse et Régulation des Processus Industriels. Tome 1. Régulation Continue.* Technip, Paris, 1993.

R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, (5):1190–1208, 1995.

B. Carnahan, H.A. Luther, and J.O. Wilkes. *Applied Numerical Methods*. Wiley, New York, 1969.

C.T. Chen. *One-dimensional Digital Signal Processing*. Marcel Dekker, New York, 1979.

C.T. Chen. *Analog and Digital Control System Design: Transfer-Function, State-Space, and Algebraic Methods*. Harcourt Brace Jovanovich College, Fort Worth, 1993.

D.R. Coughanowr and L.B. Koppel. *Process Systems Analysis and Control*. McGraw-Hill, Auckland, 1985.

R. Fletcher. *Practical Methods of Optimization*. Wiley, Chichester, 1991.

P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic Press, London, 1981.

G.C. Goodwin and K.S. Sin. *Adaptive Filtering, Prediction and Control.* Prentice Hall, Englewood Cliffs, 1984.

D.M. Himmelblau and K.B. Bischoff. *Process Analysis and Simulation.* Wiley, New York, 1968.

R. Isermann. *Digital Control Systems*, volume I. Fundamentals Deterministic Control. Springer-Verlag, 2nd edition, 1991a.

R. Isermann. *Digital Control Systems*, volume II. Stochastic Control, Multivariable Control, Adaptive Control, Applications. Springer-Verlag, 2nd edition, 1991b.

T. Kailath. *Linear Systems Theory.* Prentice Hall, Englewood Cliffs, New Jersey, 1980.

H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems.* Wiley-Interscience, New York, 1972.

O. Levenspiel, editor. *Chemical Reaction Engineering.* Wiley, 3rd edition, 1999.

W.S. Levine, editor. *The Control Handbook.* CRC Press, Boca Raton, Florida, 1996.

C.F. Lin. *Advanced Control Systems Design.* Prentice Hall, Englewood Cliffs, New Jersey, 1994.

D.G. Luenberger. *Introduction to Dynamic Systems. Theory, Models and Applications.* Wiley, New York, 1979.

W. L. Luyben. *Process Modeling, Simulation, and Control for Chemical Engineers.* McGraw-Hill, New York, 1990.

T.E. Marlin. *Process Control. Designing Processes and Control Systems for Dynamic Performance.* McGraw-Hill, Boston, 2000.

R.H. Middleton and G.C. Goodwin. *Digital Control and Estimation.* Prentice Hall, Englewood Cliffs, 1990.

K. Ogata. *Discrete-Time Control Systems.* Prentice Hall, Englewood Cliffs, New Jersey, 1987.

K. Ogata. *Modern Control Engineering.* Prentice Hall, Englewood Cliffs, New Jersey, 1997.

R.H. Perry. *Perry's Chemical Engineers' Handbook.* McGraw-Hill, New York, 6th edition, 1973.

W.H. Ray and B.A. Ogunnaike. *Process Dynamics, Modeling and Control.* Oxford University Press, 1994.

W.H. Ray and J. Szekely. *Process Optimization with Applications in Metallurgy and Chemical Engineering.* Wiley, New York, 1973.

K. Schittkowski. NLPQL: A Fortran subroutine solving constrained nonlinear programming problems. *Ann. Oper. Res.*, 5:485–500, 1985.

D.E. Seborg, T.F. Edgar, and D.A. Mellichamp. *Process Dynamics and Control.* Wiley, New York, 1989.

S.M. Shinners. *Modern Control System Theory and Design.* Wiley, New York, 1992.

F.G. Shinskey. *Process Control Systems.* McGraw-Hill, New York, 1979.

N.K. Sinha and G.P. Rao, editors. *Identification of Continuous Time Systems. Methodology and Computer Identification.* Kluwer Academic Publishers, Dordrecht, 1991.

G. Stephanopoulos. *Chemical Process Control, an Introduction to Theory and Practice.* Prentice Hall, Englewood Cliffs, New Jersey, 1984.

H. Unbehauen and G.P. Rao. *Identification of Continuous Time Systems.* North Holland, Amsterdam, 1987.

J. Villermaux. *Génie de la Réaction Chimique.* Lavoisier, Paris, 1982.

E. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data.* Communications and Control Engineering. Springer-Verlag, London, 1997.

K. Watanabe. *Adaptive Estimation and Control.* Prentice Hall, London, 1992.

W.A. Wolovich. *Automatic Control Systems, Basic Analysis and Design.* Holt, Rinehart and Winston, New York, 1994.

P.C. Young. Parameter estimation for continuous time models. A survey. *Automatica*, 17:23–39, 1981.

C. Zhu, R.H. Byrd, P. Lu, and J. Nocedal. L-BFGS-B: a limited memory FORTRAN code for solving bound constrained optimization problems. Technical report, NAM-11, EECS Department, Northwestern University, 1994.

Chapter 2

Linear Feedback Control

The by far most used control method in industry is the proportional-integral-derivative or PID controller. It is currently claimed that 90 to 95% of industrial problems can be solved by this type of controller, which is easily available as an electronic module. It allies an apparent simplicity of understanding and a generally satisfactory performance. It is based on a quasi-natural principle which consists of acting on the process according to the error between the set point and the measured output. Indeed, along the chapters of this first part, it will appear that numerous variants of PID exist and that improvements can often be brought either by better tuning or by a different configuration.

2.1 Design of a Feedback Loop

2.1.1 Block Diagram of the Feedback Loop

Feedback control consists of a reinjection of the output in a loop (Fig. 2.2). The output response y or controlled variable is used to act on the control variable (or manipulated input) u in order to make the difference $(y_r - y)$ between the desired or reference set point y_r and the output y as small as possible for any value of any disturbance d. The output y is linked to the set point y_r by a system which forces the output to follow the set point.

If a fixed value of the set point is imposed, the system is said to be regulating or in regulation mode; if the set point is variable (following a trajectory), the system is said to be tracking the set point or in tracking mode or subjected to a servo-mechanism. The trajectory tracking is often met, e.g. in the case of batch reactors in fine chemistry, a temperature or feed profile is imposed or, in the case of a gas chromatograph, a temperature profile is imposed on the oven temperature.

Open-Loop System:
The vocabulary open or closed loop, comes from the electricity domain. Thus, an electrical circuit must be imagined, open when it is an open loop, closed for a closed loop.

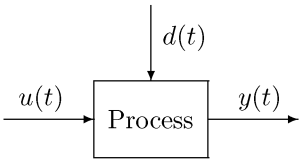


Figure 2.1: Process representation in open loop

In the open loop (Fig. 2.1), the output value is not used to correct the error. The open loop can work (theoretically) only if the process model is perfect and in the absence of disturbances: in practice, many phenomena such as measurement errors, noise and disturbances are superposed so that the use of the open loop is to be proscribed. However, feedforward control (Sect. 6.6) is a special open-loop design to counterbalance the measured disturbances; in many cases, it is coupled with feedback control.

Closed-Loop System:

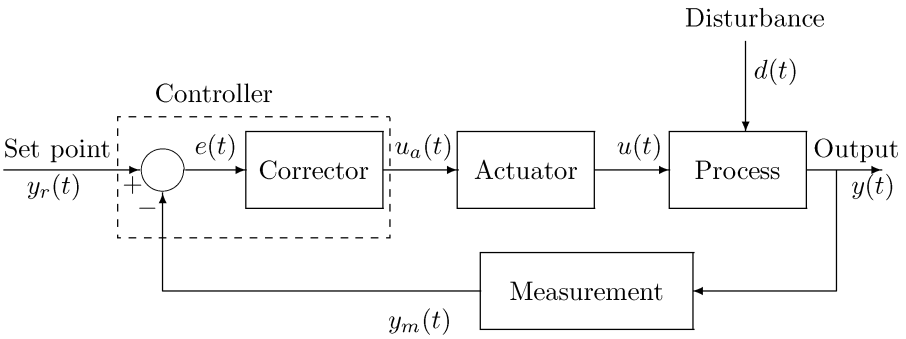


Figure 2.2: Representation of a closed-loop process

The process of Fig. 2.2 presents an output y, a disturbance d and a control variable u. In general, the shape of disturbance is unpredictable and the objective is to maintain the output y as close as possible to the desired set point y_r for any disturbance. A control possibility is to use a feedback realized by a closed loop (Fig. 2.2):

- The output is measured using a given measurement device; the value indicated by the sensor is y_m.
- This value is compared to the set point y_r, giving the difference [set point − measurement] to produce the error $e = y_r - y_m$.
- The value of this difference is provided to the main corrector, the function of which is to modify the value of the control variable u in order to reduce the error e. The corrector does not operate directly, but through an actuator (valve, transducer ...) to which it gives a value u_a.

Important remark: one acts on the actuator and modifies the control variable u only after having noted the effect of the disturbance on the output. The set of

the comparator and the corrector constitutes the control system and is called a controller which can perform regulation actions as well as tracking.

2.1.2 General Types of Controllers

A controller can take very different forms. In reality, it represents a control strategy, that is to say, a set of rules providing a value of the control action when the output deviates from the set point. A controller can thus be constituted by an equation or an algorithm.

In this first stage, only simple conventional controllers are considered.

Proportional (P) Controller

The operating output of the proportional controller is proportional to the error

$$u_a(t) = K_c\, e(t) + u_{ab} \tag{2.1}$$

where K_c is the proportional gain of the controller.
u_{ab} is the bias signal of the actuator (= operating signal when $e(t) = 0$), adjusted so that the output coincides with the desired output at steady state.

The proportional controller is characterized by the proportional gain K_c, sometimes by the proportional band PB defined by

$$PB = \frac{100}{K_c} \tag{2.2}$$

in the case of a dimensionless gain. In general, the proportional band takes values between 1 and 500; it represents the domain of error variation so that the operating signal covers all its domain. The higher the gain, the smaller the proportional band, and the more sensitive the controller. The sign of gain K_c can be positive or negative. K_c can be expressed with respect to the dimensions of output $u_a(t)$ and input $e(t)$ signals, or dimensionless according to the case.

A controller can saturate when its output $u_a(t)$ reaches a maximum $u_{a,\max}$ or minimum $u_{a,\min}$ value.

The controller transfer function is simply equal to the controller gain

$$G_c(s) = K_c \tag{2.3}$$

In the following, it will be noticed that the proportional controller presents the drawback to creating a deviation of the output with respect to the set point.

Proportional-Integral (PI) Controller

The operating output of the PI controller is proportional to the weighted sum of the magnitude and of the integral of the error

$$u_a(t) = K_c \left(e(t) + \frac{1}{\tau_I} \int_0^t e(x)\, dx \right) + u_{ab} \tag{2.4}$$

For chemical processes, the integral time constant is often around $0.1 \leq \tau_I \leq 60$ min.

The integral action tends to modify the controller output $u_a(t)$ as long as an error exists in the process output; thus an integral controller can only modify small errors. The transfer function of the PI controller is equal to

$$G_c(s) = K_c \left(1 + \frac{1}{\tau_I s}\right) \tag{2.5}$$

Philosophy: the integral action takes into account (integrates) the past.

Compared to the proportional controller, the PI controller presents the advantage of eliminating the deviation between the output and the set point owing to the integral action. However, this controller can produce oscillatory responses and diminishes the closed-loop system stability. Furthermore, the integral action can become undesirable when there is saturation: the controller acts at its maximum level and nevertheless the error persists; the phenomenon is called windup. In this case, the integral term increases largely, possibly without limitation, and it is necessary to stop the integral action. An anti-windup device must be incorporated into PI controllers (see Sect. 4.6.4).

Ideal Proportional-Derivative (PD) Controller

The operating output of the ideal PD controller is proportional to the weighted sum of the magnitude and the time rate of change of the error

$$u_a(t) = K_c \left(e(t) + \tau_D \frac{de(t)}{dt}\right) + u_{ab} \tag{2.6}$$

The derivative action is intended to anticipate future errors. The transfer function of the ideal PD controller is equal to

$$G_c(s) = K_c \left(1 + \tau_D s\right) \tag{2.7}$$

This controller is theoretical because the numerator degree of the controller transfer function $G_c(s)$ is larger than the denominator degree; consequently, it is physically unrealizable. In practice, the previous derivative action is replaced by the ratio of two first-order polynomials presenting close characteristics for low and medium frequencies. Furthermore, an integral action is always added. The derivative action has a stabilizing influence on the controlled process.
Philosophy: the derivative action takes into account (anticipates) the future.

Ideal Proportional-Integral-Derivative (PID) Controller

This type of controller is the most often used, however, in a slightly different form from the ideal one which will be first presented. The operating output of the ideal PID controller is proportional to the weighted sum of the magnitude, the integral and the time rate of change of the error

$$u_a(t) = K_c \left(e(t) + \tau_D \frac{de(t)}{dt} + \frac{1}{\tau_I} \int_0^t e(x)\, dx\right) + u_{ab} \tag{2.8}$$

The transfer function of the PID controller is equal to

$$G_c(s) = K_c \left(1 + \tau_D\, s + \frac{1}{\tau_I\, s}\right) \tag{2.9}$$

The previous remark on the physical unrealizability of the derivative action is still valid.
Philosophy: owing to the derivative action, the PID controller takes into account (anticipates) the future, and owing to the integral action, the PID takes into account (integrates) the past.
Remark 1:
The previous theoretical controller is, in practice, replaced by a real PID controller of the following transfer function

$$G_c(s) = K_c \left(\frac{\tau_I s + 1}{\tau_I s}\right) \left(\frac{\tau_D s + 1}{\beta\, \tau_D s + 1}\right) \tag{2.10}$$

which is physically realizable.
Remark 2:
It is often preferred to operate the PID controller by making the derivative action act no more on the error coming from the comparator but on the measured output, under the theoretical form

$$u_a(t) = K_c \left(e(t) + \frac{1}{\tau_I} \int_0^t e(x)\, dx - \tau_D\, \frac{dy_m}{dt}\right) \tag{2.11}$$

or practically

$$U_a(s) = K_c \left(1 + \frac{1}{\tau_I\, s}\right) E(s) - \left(\frac{K_c\, \tau_D s}{\frac{\tau_D}{N} s + 1}\right) Y_m(s) \tag{2.12}$$

This method of taking into account the derivative action allows to avoid brutal changes of the controller output due to the error signal variation.

2.1.3 Sensors

This point may seem simple; indeed, in a real process, it is an essential element. Without good measurement, it is hopeless to control the process well. The sensor itself and the information transmission chain given by the sensor are concerned. The common sensors that are met on chemical processes are:

- Temperature sensors: thermocouples, platinum resistance probes, pyrometers. Temperature sensors can be modelled from the response they give to a temperature step according to a first- or second-order model, sometimes with a time delay.
- Pressure sensors: classical manometers using bellows, Bourdon tube, membrane or electronic ones using strain gauges (semiconductors whose resistance changes under strain). Diaphragm pressure sensors use detection of the diaphragm position by measurement of electrical capacitance. They are often represented by a second-order model.

- Flow rate sensors: for gas flow such as thermal mass flow meters (based on the thermal conductivity of gases, the gas flow inducing a temperature variation in a capillary tube), variable area flow meters (displacement of a float in a conical vertical tube); for liquid flow such as turbine flow meters (rotation of a turbine), depression flow meters as venturi-type flow meters (the flow rate is proportional to the square root of the pressure drop), vortex flow meters (measurement of the frequency of vortex shedding due to the presence of an unstreamlined obstacle), electromagnetic flow meters (for electrically conducting fluids), sonic flow meters, Coriolis effect flow meters. Flow rate sensors have very fast dynamics and are often modelled by an equation of the form
$$\text{flow rate} = a\sqrt{\Delta P}$$
where the proportionality constant a is dependent on the sensor and ΔP is the pressure drop between the section restriction point and the outlet. These signals are often noisy because of flow fluctuations and should be filtered before being used by the controller.
- Level sensors: floats (lighter than the fluid), displacement (measurement of the apparent weight of a half-submerged cylinder) through a pressure difference measurement, conductivity probes indicating liquid presence, capacitance detectors for level variations.
- Composition sensors: potentiometers (chemical potential measurement of an ion by means of a specific electrode), conductimeters (measurement of a solution conductivity), chromatographs (separation of liquids or gases), spectrometers (visible, UV, infra-red, ...). Among these, chromatographs pose a particular and very important problem in practice: the information provided by these apparatus arrives a long time after the sampling and thus there exists a large time delay that must be included in the model. This time delay can be the cause of a lack of mastering or imperfect mastering of the process control.

In the absence of a measurement concerning a given variable, if a model of the process is available, it is possible to realize a state observer called a software sensor. This latter will use other available measurements to estimate the value of the unmeasured variable, it is comparable to an indirect measurement. The linear Kalman filter or nonlinear (extended Kalman filter) is often used for this purpose (see Sect. 11.1.2 and 18.4.3). Chemical composition estimations are particularly concerned by this type of sensor.

The transmitter is the intermediary between the sensitive element of the sensor and the controller. It is a simple converter which is then considered as a simple gain, ratio of the difference of the output signal (often transmitted in the range 4-20 mA) over the difference of the input signal given by the sensor.

The set sensor-transmitter can be considered as a global measurement device.

2.1.4 Transmission Lines

Traditionally, transmission lines were pneumatic. Nowadays, more and more often they are electrical lines. In general, their dynamic influence on the process is neglected except in the case of very fast phenomena, which is not very common in chemical engineering.

2.1.5 Actuators

Actuators (Considine, 1999) constitute material elements which allow action by means of the control loop on the process. For example, as a flow rate actuator, a very common element is the pneumatic valve, operating as indicated by its name with pressurized air. It could as well be mechanically operated by a dc or a stepping motor (Fig. 2.3).

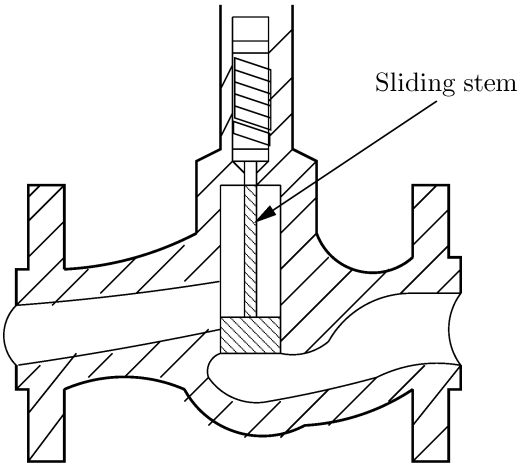


Figure 2.3: Scheme of a typical sliding stem valve

A valve is designed to be in position, either completely open (fail open, or air-to-open) or completely closed (fail closed, or air-to-close) when the air pressure is not ensured, which can happen in the case of an incident in the process. Consider the case of a valve closed in the absence of pressure: when the air pressure increases on the diaphragm, the spring is compressed and the valve plug pulls out from its seat, thus increasing the passage section for the fluid, hence the flow rate. The inverse type of valve (open in the absence of pressure) exists where the pressure increase makes the valve plug go down (either because of the position of the air inlet with respect to the diaphragm or because of the disposal of the seat and the valve plug) and thus the cross-section decreases. The choice of valves is generally made by taking into account safety rules. There also exists motorized valves: rotating valves (butterfly, ball). In general, the valve dynamics is fast. It must not be forgotten that the valve introduces a pressure drop in the pipe. With respect to control, a valve should

not be operated too close to its limits, either completely open or completely closed, where its behaviour will be neither reproducible nor easily controllable. Frequently, a valve has a highly nonlinear behaviour on all its operating range and it is necessary to linearize it piecewise and use the constructed table for the control law.

For liquids, the flow rate Q depends on the square root of the pressure drop ΔP_v caused by the valve according to

$$Q = C_v \sqrt{\frac{\Delta P}{d}} \tag{2.13}$$

where d is the fluid density (with respect to water) and C_v is a flow rate coefficient such that the ratio (C_v/D_v^2) is approximately constant for liquids for a given type of valve, D_v being the nominal valve diameter. Viscosity corrections are required for C_v in the case of viscous liquids.

For gases or vapours, when the flow is subsonic, the volume gas flow rate is

$$Q = 0.92\, C_f\, C_v\, P_{up}\, (Y - 0.148\, Y^3)\, \frac{1}{\sqrt{d_g\, T_{up}}} \tag{2.14}$$

where Q is given in $\mathrm{m}^3.\mathrm{s}^{-1}$ at 15 °C under 1 normal atm, P_{up} is the upstream pressure (in Pa), T_{up} the upstream temperature, Y the dimensionless expansion factor and d_g the gas density (with respect to air). C_f is a dimensionless factor which depends on the type of fittings of the valve and ranges from 0.80 to 0.98. Y is equal to

$$Y = \frac{1.63}{C_f} \sqrt{\frac{\Delta P}{P_{up}}} \tag{2.15}$$

When $Y < 1.5$, the flow is subsonic; when $Y \geq 1.5$, the flow is sonic, i.e. choked.
When the flow is sonic, the volume gas flow rate is

$$Q = 0.92\, C_f\, C_v\, P_{up}\, \frac{1}{\sqrt{d_g\, T_{up}}} \tag{2.16}$$

The ratio q of the real flow rate Q to the maximum flow rate Q_m

$$q = \frac{Q}{Q_m} \tag{2.17}$$

can depend on the aperture degree x of the valve in several ways (Fig. 2.4). Denoting the sensibility as $\sigma = dq/dx$, the latter can be constant (linear behaviour: case 1), or increase with x (case 2), or decrease with x (case 3), or increase then decrease (case 4) (Midoux, 1985). Often, three main types of valves are distinguished (Thomas, 1999): with linear characteristics, with butterfly characteristics and with equal percentage characteristics. Denoting the position of the valve positioner by x, the equations for the valve constant are

respectively

$$\begin{array}{ll} \text{Linear:} & C_v = C_{vs}\, x \\ \text{Butterfly:} & C_v = C_{vs}\,(1 - \cos(\frac{\pi}{2}x)) \\ \text{Equal percentage:} & C_v = C_{vs}\, R_v^{x-1} \end{array} \tag{2.18}$$

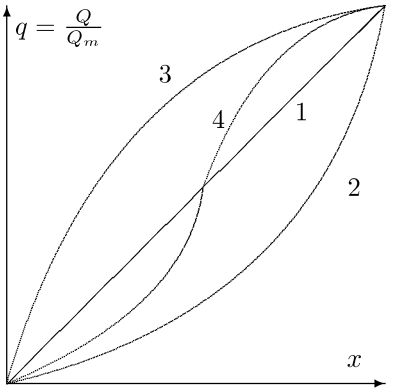


Figure 2.4: Influence of the aperture degree of a valve on its flow rate

2.2 Block Diagrams, Signal-Flow Graphs, Calculation Rules

The study of feedback control for single-input single-output processes is performed in this chapter using Laplace transfer functions. It would be possible to do the same technical realizations and their theoretical study based on state-space modelling. On the other hand, the theoretical discussion and mathematical tools would be completely different. A sketch of the state-space study nevertheless will be presented.

When specialized packages for solving control problems (e.g. MATLAB©) are used, it is very easy to find the state-space model equivalent to a transfer function. Furthermore, it is possible to set blocks in series or in parallel, to do feedback loops, either with transfer functions or in state space, and then to realize a complete block diagram in view of a simulation. However, an important difference exists between both approaches. When the studied system becomes complicated, the numerical solving based on transfer functions gives worse and maybe erroneous results, compared to the complete state-space solving. The reason is in the far more direct approach of the phenomena in state space and their direct numerical solving.

Given a block diagram in which each block represents a transfer function, the output of any block must be calculated with respect to the input of any other block. Beyond the blocks of transfer functions, the block diagram uses

summators, which do the algebraic addition of inlet signals, and signal dividers, which separate a signal into two or several signals of same intensity. Most of the common cases are represented in Figs. 2.5 to 2.11 and results are given in the transfer function case and in the state-space case.

Calculation rules with Laplace transform

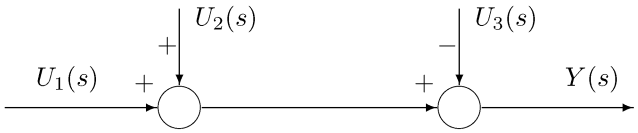


Figure 2.5: Block scheme number 1

For block scheme number 1 (Fig. 2.5), which contains two summators, the Laplace transform equation is

$$Y(s) = U_1(s) + U_2(s) - U_3(s) \tag{2.19}$$

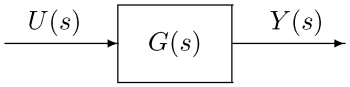


Figure 2.6: Block scheme number 2

For block scheme number 2 (Fig. 2.6), which contains only one transfer function, the Laplace transform equation is

$$Y(s) = GU(s) \tag{2.20}$$

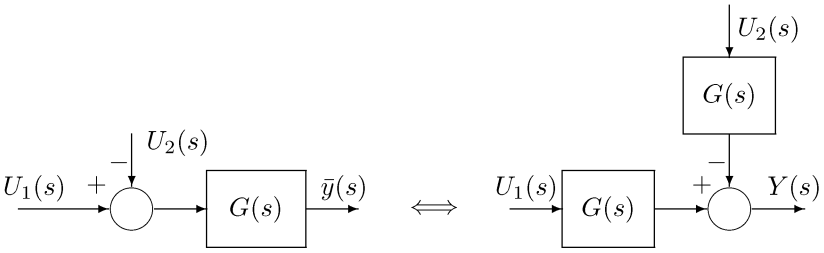


Figure 2.7: Block scheme number 3 under two equivalent representations

For block scheme number 3 (Fig. 2.7), which contains one transfer function and a summator, the Laplace transform equation is

$$Y(s) = G\left(U_1(s) - U_2(s)\right) \tag{2.21}$$

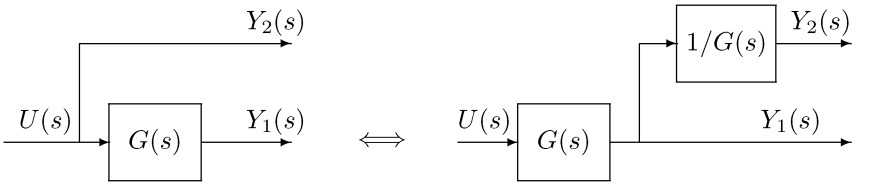


Figure 2.8: Block scheme number 4 under two equivalent representations

For block scheme number 4 (Fig. 2.8), which contains one transfer function and a signal divider, the Laplace transform equations are

$$Y_1(s) = G\,U(s) \qquad ; \quad Y_2(s) = U(s) \tag{2.22}$$

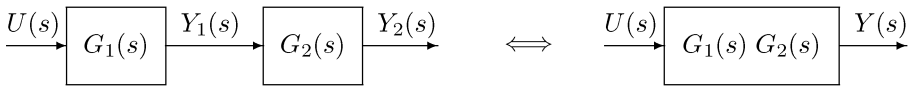


Figure 2.9: Block scheme number 5 under two equivalent representations

For block scheme number 5 (Fig. 2.9), which contains two transfer functions in series, the Laplace transform equation is

$$Y(s) = G_1\,G_2\,U(s) \tag{2.23}$$

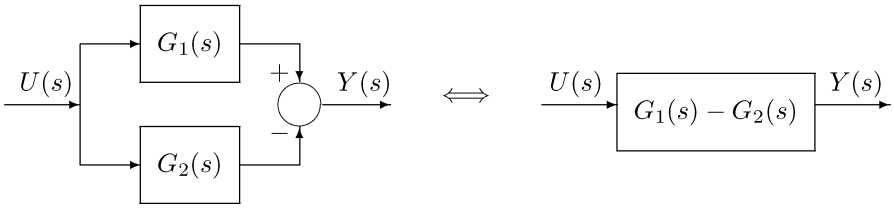


Figure 2.10: Block scheme number 6 under two equivalent representations

For block scheme number 6 (Fig. 2.10), which contains two transfer functions in parallel and a summator, the Laplace transform equation is

$$Y(s) = (G_1 - G_2)\,U(s) \tag{2.24}$$

For block scheme number 7 (Fig. 2.11), which contains two transfer functions in a feedback loop containing a summator with the feedback of sign ϵ ($\epsilon = +1$ for a positive feedback, $\epsilon = -1$ for a negative feedback), the Laplace transform equation is

$$Y(s) = \frac{G_1}{1 + \epsilon\,G1\,G_2}\,U(s) \tag{2.25}$$

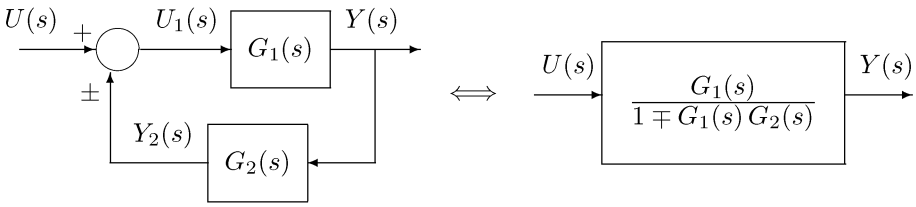


Figure 2.11: Block scheme number 7 under two equivalent representations

Calculation rules in state space

In state space, signals are directly considered with respect to the time variable and each system or block number i is represented by the set of matrices (A_i, B_i, C_i, D_i). Recall that if a system can be represented by a strictly proper transfer function, the matrix D_i is zero and this is the case of most physical systems. Equations are given in the case of single-input single-output systems.

Block scheme number 1

$$y(t) = u_1(t) + u_2(t) - u_3(t) \tag{2.26}$$

Block scheme number 2

$$\begin{cases} \dot{\boldsymbol{x}}(t) = & \boldsymbol{A}\,\boldsymbol{x}(t) + \boldsymbol{B}\,u(t) \\ y(t) = & \boldsymbol{C}\,\boldsymbol{x}(t) + \boldsymbol{D}\,u(t) \end{cases} \tag{2.27}$$

Block scheme number 3

$$\begin{cases} \dot{\boldsymbol{x}}(t) = & \boldsymbol{A}\,\boldsymbol{x}(t) + \boldsymbol{B}\,(u_1(t) - u_2(t)) \\ y(t) = & \boldsymbol{C}\,\boldsymbol{x}(t) + \boldsymbol{D}\,(u_1(t) - u_2(t)) \end{cases} \tag{2.28}$$

Block scheme number 4

$$\begin{cases} \dot{\boldsymbol{x}}(t) = & \boldsymbol{A}\,\boldsymbol{x}(t) + \boldsymbol{B}\,u(t) \\ y_1(t) = & \boldsymbol{C}\,\boldsymbol{x}(t) + \boldsymbol{D}\,u(t) \\ y_2(t) = & u(t) \end{cases} \tag{2.29}$$

Block scheme number 5, two systems in series: equations for each block are the following

$$\begin{cases} \dot{\boldsymbol{x}}_1(t) = & \boldsymbol{A}_1\,\boldsymbol{x}_1(t) + \boldsymbol{B}_1\,u(t) \\ y_1(t) = & \boldsymbol{C}_1\,\boldsymbol{x}_1(t) + \boldsymbol{D}_1\,u(t) \\ \dot{\boldsymbol{x}}_2(t) = & \boldsymbol{A}_2\,\boldsymbol{x}_2(t) + \boldsymbol{B}_2\,y_1(t) \\ y(t) = & \boldsymbol{C}_2\,\boldsymbol{x}_2(t) + \boldsymbol{D}_2\,y_1(t) \end{cases} \tag{2.30}$$

Defining the global state vector, union of both state vectors:

$$\boldsymbol{x}(t) = \begin{bmatrix} \boldsymbol{x}_1(t) \\ \boldsymbol{x}_2(t) \end{bmatrix} \tag{2.31}$$

one obtains for two systems in series

$$\left\{\begin{array}{ll} \dot{\boldsymbol{x}}(t) = & \left[\begin{array}{c} \dot{\boldsymbol{x}}_1(t) \\ \dot{\boldsymbol{x}}_2(t) \end{array}\right] = \left[\begin{array}{cc} \boldsymbol{A}_1 & 0 \\ \boldsymbol{B}_2\,\boldsymbol{C}_1 & \boldsymbol{A}_2 \end{array}\right] \boldsymbol{x}(t) + \left[\begin{array}{c} \boldsymbol{B}_1 \\ \boldsymbol{B}_2\,\boldsymbol{D}_1 \end{array}\right] u(t) \\ y(t) = & \left[\begin{array}{cc} \boldsymbol{D}_2\,\boldsymbol{C}_1 & \boldsymbol{C}_2 \end{array}\right] \boldsymbol{x}(t) + \boldsymbol{D}_2\,\boldsymbol{D}_1\,u(t) \end{array}\right. \tag{2.32}$$

Block scheme number 6, two systems in parallel

$$\left\{\begin{array}{ll} \dot{\boldsymbol{x}}(t) = & \left[\begin{array}{c} \dot{\boldsymbol{x}}_1(t) \\ \dot{\boldsymbol{x}}_2(t) \end{array}\right] = \left[\begin{array}{cc} \boldsymbol{A}_1 & 0 \\ 0 & \boldsymbol{A}_2 \end{array}\right] \boldsymbol{x}(t) + \left[\begin{array}{c} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{array}\right] u(t) \\ y(t) = & \left[\begin{array}{cc} \boldsymbol{C}_1 & -\boldsymbol{C}_2 \end{array}\right] \boldsymbol{x}(t) + (\boldsymbol{D}_1 - \boldsymbol{D}_2)\,u(t) \end{array}\right. \tag{2.33}$$

Block scheme number 7, feedback loop of sign ϵ:
the general case where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are not zero is first treated.
The basic equations are the following

$$\left\{\begin{array}{ll} \dot{\boldsymbol{x}}_1(t) = & \boldsymbol{A}_1\,\boldsymbol{x}_1(t) + \boldsymbol{B}_1\,u_1(t) \\ y(t) = & \boldsymbol{C}_1\,\boldsymbol{x}_1(t) + \boldsymbol{D}_1\,u_1(t) \\ \dot{\boldsymbol{x}}_2(t) = & \boldsymbol{A}_2\,\boldsymbol{x}_2(t) + \boldsymbol{B}_2\,y(t) \\ y_2(t) = & \boldsymbol{C}_2\,\boldsymbol{x}_2(t) + \boldsymbol{D}_2\,y(t) \\ u_1(t) = & u(t) + \epsilon\,y_2(t) \end{array}\right. \tag{2.34}$$

Eliminating internal variables $u_1(t)$ and $y_2(t)$, one obtains

$$\begin{array}{l} \dot{\boldsymbol{x}}(t) = \left[\begin{array}{c} \dot{\boldsymbol{x}}_1(t) \\ \dot{\boldsymbol{x}}_2(t) \end{array}\right] = \\ \left[\begin{array}{ll} \boldsymbol{A}_1 + \epsilon \boldsymbol{B}_1 \boldsymbol{D}_2 \boldsymbol{C}_1\,[\boldsymbol{I} - \epsilon \boldsymbol{D}_1 \boldsymbol{D}_2]^{-1} & \epsilon \boldsymbol{B}_1 \boldsymbol{C}_2 + \boldsymbol{B}_1 \boldsymbol{D}_2 \boldsymbol{D}_1 \boldsymbol{C}_2\,[\boldsymbol{I} - \epsilon \boldsymbol{D}_1 \boldsymbol{D}_2]^{-1} \\ \boldsymbol{B}_2 \boldsymbol{C}_1\,[\boldsymbol{I} - \epsilon \boldsymbol{D}_1 \boldsymbol{D}_2]^{-1} & \boldsymbol{A}_2 + \epsilon \boldsymbol{B}_2 \boldsymbol{D}_1 \boldsymbol{C}_2\,[\boldsymbol{I} - \epsilon \boldsymbol{D}_1 \boldsymbol{D}_2]^{-1} \end{array}\right] \boldsymbol{x}(t) \\ + \left[\begin{array}{l} \boldsymbol{B}_1 + \epsilon\,\boldsymbol{B}_1\,\boldsymbol{D}_2\,\boldsymbol{D}_1\,[\boldsymbol{I} - \epsilon\,\boldsymbol{D}_1\,\boldsymbol{D}_2]^{-1} \\ \boldsymbol{B}_2\,\boldsymbol{D}_1\,[\boldsymbol{I} - \epsilon\,\boldsymbol{D}_1\,\boldsymbol{D}_2]^{-1} \end{array}\right] u(t) \\ y(t) = \left[\begin{array}{l} \boldsymbol{C}_1\,[\boldsymbol{I} - \epsilon\,\boldsymbol{D}_1\,\boldsymbol{D}_2]^{-1} \\ \epsilon\,\boldsymbol{D}_1\,\boldsymbol{C}_2\,[\boldsymbol{I} - \epsilon\,\boldsymbol{D}_1\,\boldsymbol{D}_2]^{-1} \end{array}\right] \boldsymbol{x}(t) + \boldsymbol{D}_1\,[\boldsymbol{I} - \epsilon\,\boldsymbol{D}_1\,\boldsymbol{D}_2]^{-1}\,u(t) \end{array} \tag{2.35}$$

When both transfer functions are strictly proper, and matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ zero, equations can be simplified as

$$\left\{\begin{array}{ll} \dot{\boldsymbol{x}}(t) = & \left[\begin{array}{c} \dot{\boldsymbol{x}}_1(t) \\ \dot{\boldsymbol{x}}_2(t) \end{array}\right] = \left[\begin{array}{cc} \boldsymbol{A}_1 & \epsilon\,\boldsymbol{B}_1\,\boldsymbol{C}_2 \\ \boldsymbol{B}_2\,\boldsymbol{C}_1 & \boldsymbol{A}_2 \end{array}\right] \boldsymbol{x}(t) + \left[\begin{array}{c} \boldsymbol{B}_1 \\ 0 \end{array}\right] u(t) \\ y(t) = & \left[\begin{array}{cc} \boldsymbol{C}_1 & 0 \end{array}\right] \boldsymbol{x}(t) \end{array}\right. \tag{2.36}$$

Mason Formula and Signal-Flow Graphs
The Mason formula allows us to quickly calculate global transfer functions for a block scheme where each block represents a transfer function. A complicated scheme such as in Fig. 2.12 is considered.

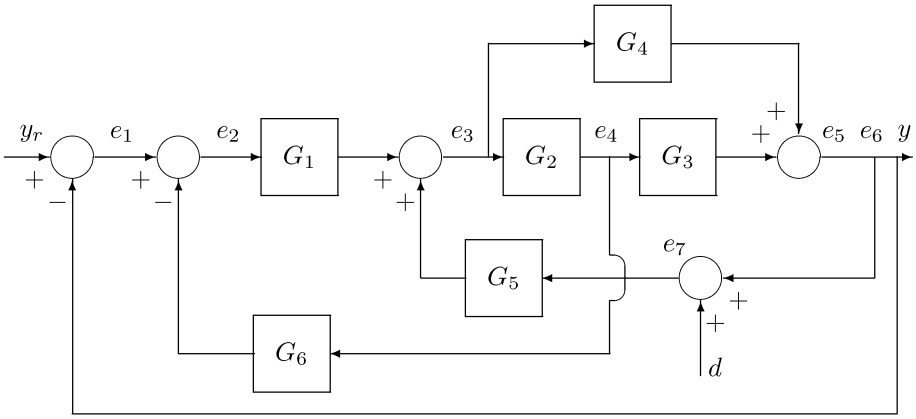


Figure 2.12: Block scheme of a closed-loop process

This block scheme has two external inputs y_r and d, and the transfer functions set point-output G_{yr} and disturbance-output G_{yd} must be respectively calculated

$$Y(s) = G_{yr}\, Y_r(s) + G_{yd}\, D(s) \tag{2.37}$$

Each block is directed from input towards output and is called unidirectional. A loop is a unidirectional path which starts and ends at a same point, and along which no point is met more than once. A loop transmittance is equal to the product of the transfer functions of the loop. When a loop includes summators, the concerned signs must be taken into account in the calculation of the loop transmittance.

Rather than directly working on the block diagram such as is currently described, it is preferable to transform this scheme into a signal-flow graph, which contains the topological information of the set of linear equations included in the block diagram. The word signal-flow (or signal flow) means a flow of fluxes or signals.

To operate, some characteristic definitions of signal-flow graphs must be added:

- A signal-flow graph is made of nodes and connecting branches (a line with an arrow).
- A node is attributed to each variable which occurs in the system. The node i represents the variable y_i for example.
- For a branch beginning in i and ending in j, the transmittance a_{ij} of the branch relates variables y_i and y_j.
- A source is a node from where only branches go out.
- A sink is a node where only branches come in.
- A path is a group of connected branches having the same direction.
- A direct path comes from a source and ends in a sink; furthermore, no node should be met more than once.

- A path transmittance is the product of the transmittances associated with the branches of this path.
- A feedback loop B_i is a path coming from a node i and ending at the same node i. Along a loop, a given node cannot be met more than once.
- A transmittance of a loop B_i is the product of the transmittances associated with the branches of this loop.
- Loops B_i and B_j are non-touching when they have no node in common.

First, let us present some simple cases that allow us to understand signal-flow graphs as well as the associated equations, which are all linear.

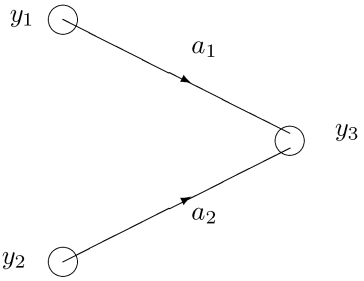


Figure 2.13: Signal-flow graph: addition

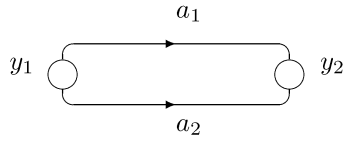


Figure 2.14: Signal-flow graph: addition

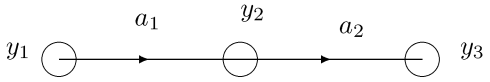


Figure 2.15: Signal-flow graph: multiplication

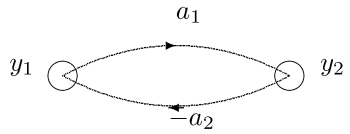


Figure 2.16: Signal-flow graph: feedback

Additions:
Graph 2.13 corresponds to the linear equation

$$y_3 = a_1\, y_1 + a_2\, y_2 \tag{2.38}$$

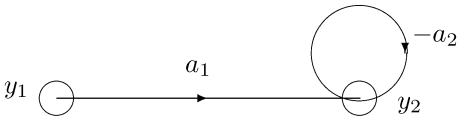


Figure 2.17: Signal-flow graph: feedback

and graph 2.14 corresponds to the linear equation

$$y_2 = (a_1 + a_2)\, y_1 \tag{2.39}$$

Multiplication:
Graph 2.15 corresponds to the linear equation

$$y_3 = a_2\, y_2 = a_1\, a_2\, y_1 \tag{2.40}$$

The transmittance of path from 1 to 3 is $a_1 a_2$.
Feedback:
Graph 2.16 corresponds to the linear equation

$$y_2 = a_1\, y_1 - a_2\, a_1\, y_2 \Longrightarrow y_2 = \frac{a_1}{1 + a_1\, a_2}\, y_1 \tag{2.41}$$

The transmittance of path from 1 to 2 is $\frac{a_1}{1+a_1\, a_2}$.
Graph 2.17 corresponds to the linear equation

$$y_2 = a_1\, y_1 - a_2\, y_2 \Longrightarrow y_2 = \frac{a_1}{1 + a_2}\, y_1 \tag{2.42}$$

Then, the Δ characteristic function of the block scheme or determinant of the signal-flow graph is defined as

$\Delta = 1$ $-\sum$ (transmittances of the loops),
$+\sum$ (products of the transmittances of all non-touching loops considered two by two),
$-\sum$ (products of the transmittances of all non-touching loops considered three by three),
...

Note that Δ is independent of the input and the output.

- A direct path from an input u_i to an output y_j is any connection of directed branches and of blocks between i and j such that no point is met more than once. The input u_i and the output y_j are connected by k direct paths each having the transmittance T_{ijk}. Let Δ_{ijk} also be the determinant of each direct path calculated according to the previous formula, giving Δ by setting equal to 0 all transmittances of the loops which touch the k-th direct path from i to j (suppress all the nodes and the branches of this direct path).

According to the Mason formula, the transfer function of the input u_i to the output y_j is equal to

$$G_{ij} = \frac{\sum_{k=1}^{k_{ij}} T_{ijk}\, \Delta_{ijk}}{\Delta} \tag{2.43}$$

with

$$Y_j(s) = G_{ij}\, U_i(s) \tag{2.44}$$

The signal-flow graph corresponding to the previous block diagram 2.12 is given in Fig. 2.18.

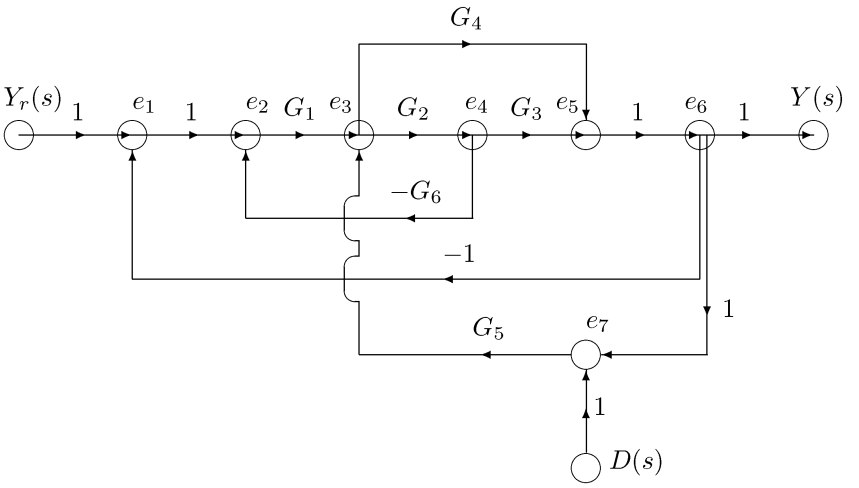


Figure 2.18: Signal-flow graph corresponding to block diagram 2.12

It might have been possible on this graph to merge E_6 and Y: variables have here been distinguished to make the sink Y clearly appear.

In this signal-flow graph, one wishes to calculates transfer functions from $Y_r(s)$ to $Y(s)$ and from $D(s)$ to $Y(s)$, respectively G_{ry} and G_{dy}. In this graph, five loops exist with respective transmittances: $G_2G_3G_5$, $-G_1G_2G_3$, $-G_1G_2G_6$, $-G_1G_4$ and G_4G_5. There exist no two by two non-touching loops. The determinant of this graph is thus equal to

$$\begin{aligned} \Delta &= 1 - (G_2G_3G_5 - G_1G_2G_3 - G_1G_2G_6 - G_1G_4 + G_4G_5) \\ &= 1 + G_1G_2G_6 + (G_2G_3 + G_4)(G_1 - G_5) \end{aligned} \tag{2.45}$$

To find the transfer function G_{ry}, two direct paths must be noticed, one is $re_1e_2e_3e_4e_5e_6y$ with transmittance $T_{ry1} = G_1G_2G_3$ for which $\Delta_{ry1} = 1$, the other one is $re_1e_2e_3e_5e_6y$ with transmittance $T_{ry2} = G_1G_4$ for which $\Delta_{ry2} = 1$. The transfer function G_{ry} is thus equal to

$$G_{ry} = \frac{G_1G_2G_3 + G_1G_4}{1 + G_1G_2G_6 + (G_2G_3 + G_4)(G_1 - G_5)} \tag{2.46}$$

To find the transfer function G_{dy}, there exist two direct paths: $de_7e_3e_4e_5e_6y$ with transmittance $T_{dy1} = G_2G_3G_5$ for which $\Delta_{dy1} = 1$, the other is $de_7e_3e_5e_6y$ with transmittance $T_{dy2} = G_4G_5$ for which $\Delta_{dy2} = 1$. The transfer function G_{dy} is thus equal to

$$G_{dy} = \frac{G_2G_3G_5 + G_4G_5}{1 + G_1G_2G_6 + (G_2G_3 + G_4)(G_1 - G_5)} \tag{2.47}$$

Globally, one obtains

$$Y(s) = G_{ry}\, Y_r(s) + G_{dy}\, D(s) \tag{2.48}$$

2.3 Dynamics of Feedback-Controlled Processes

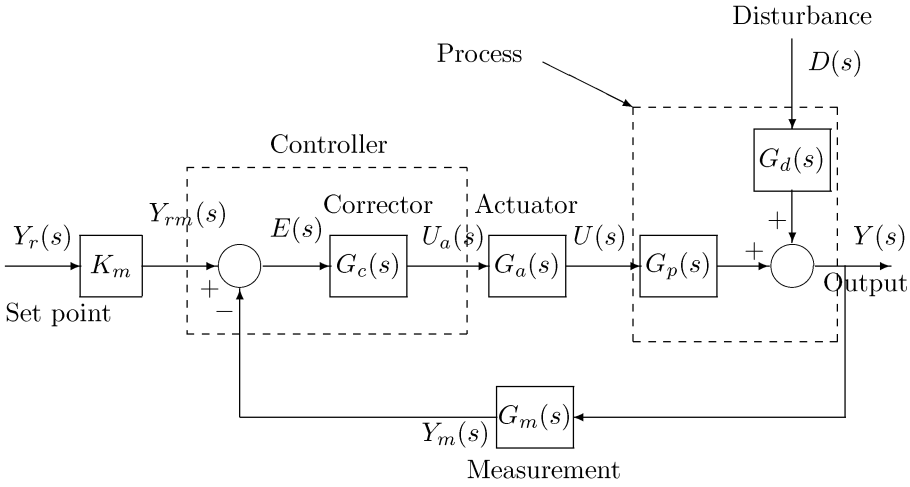


Figure 2.19: Block scheme of the closed-loop process

The block scheme of feedback control makes use of previously studied elements with respect to their general operating principle. In the block scheme of the process and control system (Fig. 2.19), independent external inputs are on one hand the set point $y_r(t)$ imposed by the user and on the other hand the disturbance $d(t)$ not mastered by the user; these inputs influence the output $y(t)$. Indeed, the process could be subjected to several disturbances. The process undergoes differently the action of the control variable $u(t)$ and of the disturbance $d(t)$, thus this corresponds to distinct transfer functions denoted respectively by $G_d(s)$ and $G_p(s)$, so that, as a Laplace transform, the output $Y(s)$ is written as

$$Y(s) = G_p(s)U(s) + G_d(s)D(s) \tag{2.49}$$

A summator will be used for the block representation. Instead of representing the block diagram in time space, it is represented as a function of the Laplace variable s.

Other transfer functions indicate the functions of different devices:

• Measurement:

$$Y_m(s) = G_m(s)Y(s) \tag{2.50}$$

It must be noted that the measured variable $y_m(t)$ generally does not have the same dimension as the corresponding output $y(t)$. For example, if the output is a temperature expressed in degrees Celsius, the variable measured by

a thermocouple is in mV, hence the steady-state gain of the transfer function has the dimension of mV/Celsius. Similarly, for any transfer function, the steady-state gain has unit dimensions. Moreover, the sensor may introduce dynamics given by the transfer function $G_m(s)$.

• Regulation:

$$U_a(s) = G_c(s)E(s) \tag{2.51}$$

with $E(s)$ being the error equal to

$$E(s) = K_m\, Y_r(s) - Y_m(s) = K_m\, Y_r(s) - G_m Y(s) \tag{2.52}$$

• Compensation of the measurement:

If the set point y_r is expressed in the same units as the output y (pressure in bar, temperature in C or K, ...), it is necessary to introduce a block to compensate the measurement (Fig. 2.19) so that the measured output y_m and the compensated set point y_{rm} have the same dimension (e.g. mA or mV), which is in general different from the output one. The gain K_m of the measurement compensation block is equal to the steady-state gain of the measurement transfer function G_m.

It is also possible to express the set point $y_r(t)$ in the same units as the measured output $y_m(t)$ and, in this case, it is not necessary anymore to compensate the measurement ($K_m = 1$).

In the case of measurement compensation, this pure gain K_m is calculated by

$$K_m = \lim_{s \to 0} G_m(s) = G_m(0) \tag{2.53}$$

so that the compensated set point is equal to

$$Y_{rm}(s) = K_m\, Y_r(s) \tag{2.54}$$

• Actuator:

$$U(s) = G_a(s)U_a(s) \tag{2.55}$$

From these equations, it is interesting to express the output $Y(s)$ with respect to the set point $Y_r(s)$ and the disturbance $D(s)$. One obtains

$$Y(s) = G_p(s)\, G_a(s)\, G_c(s)\, E(s) + G_d(s)D(s) \tag{2.56}$$

or, by expressing $E(s)$,

$$Y(s) = G_p(s)\, G_a(s)\, G_c(s)\, [K_m\, Y_r(s) - G_m(s)\, Y(s)] + G_d(s)D(s) \tag{2.57}$$

The process closed-loop response $Y(s)$ is thus equal to

$$Y(s) = \frac{G_p(s)\, G_a(s)\, G_c(s)\, K_m}{1 + G_p(s)\, G_a(s)\, G_c(s)\, G_m(s)} Y_r(s) + \frac{G_d(s)}{1 + G_p(s)\, G_a(s)\, G_c(s)\, G_m(s)} D(s) \tag{2.58}$$

The first term represents the influence of a change of the set point $Y_r(s)$ and the second term the influence of a change of disturbance $D(s)$. The closed-loop transfer function for a set point variation will be

$$G_{\text{set point}} = \frac{G_p(s)\,G_a(s)\,G_c(s)\,K_m}{1 + G_p(s)\,G_a(s)\,G_c(s)\,G_m(s)} \tag{2.59}$$

and similarly the closed-loop transfer function for a disturbance variation

$$G_{\text{disturbance}} = \frac{G_d(s)}{1 + G_p(s)\,G_a(s)\,G_c(s)\,G_m(s)} \tag{2.60}$$

The denominators of both closed-loop transfer functions are identical. These closed-loop transfer functions depend not only on the process dynamics, but also on the actuator, measurement device and the controller's own dynamics.

The application of the Mason formula would give the previous expressions. The signal-flow graph is given by Fig. 2.20.

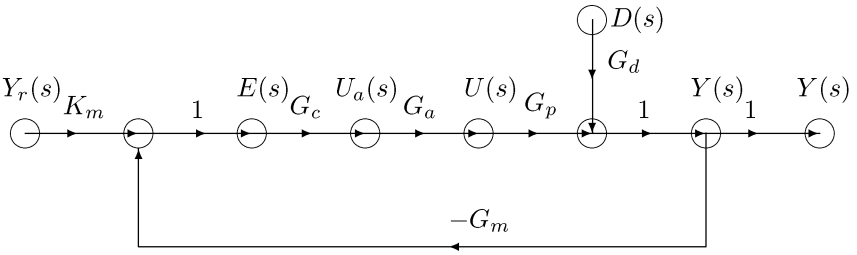


Figure 2.20: Signal-flow graph corresponding to the block diagram of the feedback control (Fig. 2.19)

In the present case, only one loop exists, and the graph determinant is equal to

$$\Delta = 1 - (-G_c G_a G_p G_m) \tag{2.61}$$

Between the set point and the output, only one direct path exists with transmittance: $T_1 = K_m G_c G_a G_p$ and determinant $\Delta_1 = 1$, so that the transfer function from the set point to the output is equal to

$$\frac{Y(s)}{Y_r(s)} = \frac{G_c G_a G_p K_m}{1 + G_c G_a G_p G_m} \tag{2.62}$$

Similarly between the disturbance and the output, only one direct path exists with transmittance: $T_1 = G_d$ and determinant $\Delta_1 = 1$, so that the transfer function from the disturbance to the output is equal to

$$\frac{Y(s)}{D(s)} = \frac{G_d}{1 + G_c G_a G_p G_m} \tag{2.63}$$

The calculation of the transfer functions can be resumed in this simple case in the following manner:

The closed-loop transfer function is equal to [product of the transfer functions met on the path between an input and an output] over [1 + the product of all transfer functions met in the loop]. So, between $D(s)$ and $Y(s)$ only G_d is met, while between $Y_r(s)$ and $Y(s)$ we meet K_m, G_c, G_a, G_p. In the loop, G_c, G_a, G_p, G_m are met. The product $G_c\,G_a\,G_p\,G_m$, which appears in the denominator of the closed-loop transfer functions, is often called open-loop transfer function, as it corresponds to the transfer function of the open loop obtained by opening the loop before the comparator as can be done for an electrical circuit (Fig. 2.21). This open-loop transfer function acts as an important role in the study of the stability of the closed-loop system.

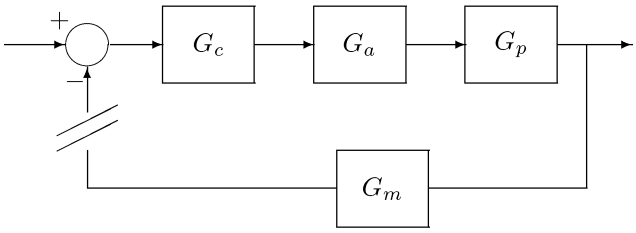


Figure 2.21: How the "open loop" must be understood by opposition to the closed loop

Two types of control problems will be studied in particular:
Regulation:
The set point is fixed ($Y_r(s) = 0$) and the process is subjected to disturbances. The control system reacts so as to maintain $y(t)$ at the set point value and tries to reject the disturbances: it is also called disturbance rejection.
Tracking:
It is assumed (in order to simplify the study) that the disturbance is constant ($D(s) = 0$) while the set point is now variable; the problem is to maintain $y(t)$ as close as possible to varying $y_r(t)$.

2.3.1 Study of Different Actions

To display the influence of different actions, only first- and second-order systems will be studied. Moreover, to simplify calculations, it will be assumed that the transfer functions of the actuator and of measurement are both equal to unity

$$G_a = 1\ ,\ G_m = 1\ ,\ K_m = 1 \tag{2.64}$$

resulting in the simplified Fig. 2.22. For these first- and second-order systems with different types of controller, responses to steps of set point or disturbance are given in Figs. 2.29, 2.30, 2.31, 2.32, and commented on in the following sections.

As an aside, to simplify, it will be assumed that the order of process transfer function G_p and the order of the transfer function G_d dealing with the distur-

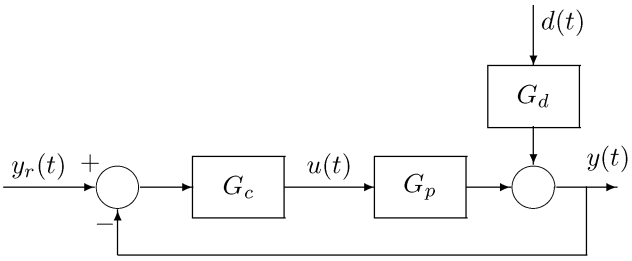


Figure 2.22: Block diagram for the study of the action of the different controllers

bance are equal (which is by no means compulsory), but that these transfer functions have different gains and time constants.

2.3.2 Influence of Proportional Action

As the controller is proportional, its transfer function is

$$G_c = K_c \tag{2.65}$$

First-Order Systems

A first-order system is described by a differential equation such as

$$\tau_p \frac{dy(t)}{dt} + y(t) = u(t) \tag{2.66}$$

where $y(t)$ is a deviation variable such that $y(0) = 0$ and $(dy/dt)_0 = 0$.
The process transfer function linking the output Laplace transform $Y(s)$ to the input Laplace transform $U(s)$ is equal to

$$G_p(s) = \frac{K_p}{\tau_p\, s + 1} \tag{2.67}$$

The transfer function for the disturbance is also assumed to be first-order

$$G_d(s) = \frac{K_d}{\tau_d\, s + 1} \tag{2.68}$$

The output $Y(s)$ for any set point $Y_r(s)$ and any disturbance $D(s)$ is thus equal to

$$Y(s) = \frac{K_p\, K_c}{\tau_p\, s + 1 + K_p\, K_c}\, Y_r(s) + \frac{K_d}{\tau_d\, s + 1}\, \frac{\tau_p\, s + 1}{\tau_p\, s + 1 + K_p\, K_c}\, D(s) \tag{2.69}$$

If we set $s \to 0$ (equivalent to an infinite time) in the transfer functions, the closed-loop transfer functions are respectively equal to the closed-loop steady-state gains (use of the final value theorem) for the set point y_r and the disturbance d

$$K'_p = \frac{K_p\, K_c}{1 + K_p\, K_c} \tag{2.70}$$

$$K'_d = \frac{K_d}{1 + K_p\, K_c} \tag{2.71}$$

The closed-loop gain K'_p is modified compared to the open-loop gain K_p; K'_p tends towards 1 when the controller gain is large. The closed-loop gain relative to the disturbance K'_d is lower than the open-loop gain K_d and tends towards 0 when the controller gain is large. The closed-loop response is still first-order with respect to set point and disturbance variations.

The open-loop time constant is τ_p; in closed loop, concerning set point variations, it is equal to

$$\tau'_p = \frac{\tau_p}{1 + K_p\, K_c} \tag{2.72}$$

thus it has decreased; the response will be faster in closed loop than in open loop.

Consider the response to a step variation of set point (tracking) or disturbance (regulation):

Tracking study:

The set point change is a step of amplitude A

$$Y_r(s) = \frac{A}{s} \tag{2.73}$$

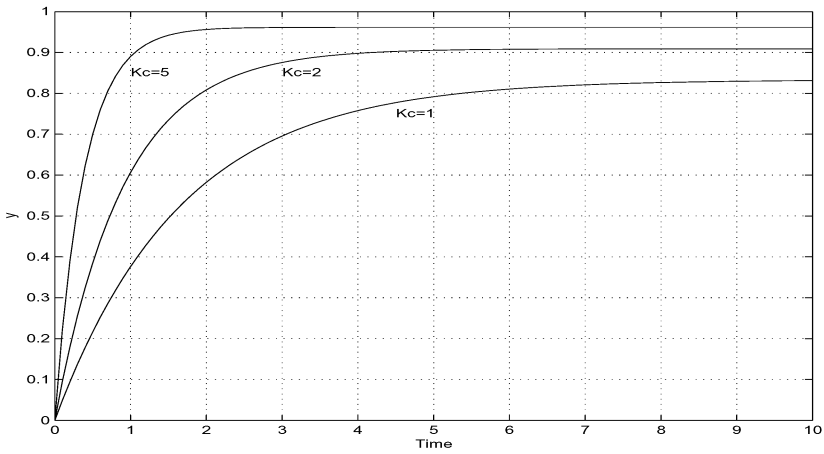


Figure 2.23: Response of a first-order system ($K_p = 5, \tau_p = 10$) to a set point unit step (proportional controller with increasing gain: $K_c = 1, 2, 5$)

The disturbance is assumed constant or zero ($D(s) = 0$). The closed-loop response (Fig. 2.23) to a set point step is then equal to

$$Y(s) = \frac{K_p\, K_c}{\tau_p\, s + 1 + K_p\, K_c} \frac{A}{s} = \frac{K'_p}{\tau'_p\, s + 1} \frac{A}{s} \tag{2.74}$$

To get the time response $y(t)$, $Y(s)$ is decomposed into a sum of rational fractions, the first corresponding to the forced response $Y_f(s)$, the second to the natural response $Y_n(s)$

$$Y(s) = A\frac{K'_p}{s} - A\frac{K'_p\,\tau'_p}{\tau'_p s + 1} = Y_f(s) + Y_n(s) \tag{2.75}$$

hence

$$\begin{array}{l} y(t) = AK'_p\left(1 - \exp(-t/\tau'_p)\right) = y_f(t) + y_n(t) \\ \text{with:} \quad y_f(t) = AK'_p \quad ; \quad y_n(t) = -AK'_p \exp(-t/\tau'_p) \end{array} \tag{2.76}$$

Figure 2.23 was obtained for a unit step of a set point. The asymptotic value of the output presents an offset with the set point; if the controller gain K_c is increased, this offset decreases (Fig. 2.23). In practice, other transfer functions must be taken into account: actuator, measurement, so that this set may not behave exactly as a first-order process and present e.g. a time delay, nonlinearities or neglected dynamics. The choice of too large gains for the proportional controller may render the closed-loop behaviour oscillatory or unstable. A high gain decreases the response time, imposing a more important demand to the actuator: the control variable $u(t)$ varies more strongly and more quickly so that it can reach its limits, it is then saturated.

Regulation study:

Recall that the transfer function for the disturbance was taken to be first-order as that of the process, but with different gain $K_d \neq K_p$ and time constant $\tau_d \neq \tau_p$.

Consider a disturbance step variation of amplitude A

$$D(s) = \frac{A}{s} \tag{2.77}$$

The set point is assumed constant (regulation). The closed-loop response (Fig. 2.31) to a step disturbance is then equal to

$$Y(s) = \frac{K_d}{\tau_d s + 1}\,\frac{\tau_p s + 1}{\tau_p s + 1 + K_p K_c}\,\frac{A}{s} \tag{2.78}$$

To get the time response $y(t)$, $Y(s)$ is decomposed into

$$\begin{array}{rcl} Y(s) & = & A\left(\dfrac{K_d}{1 + K_p\,K_c}\,\dfrac{1}{s} + \dfrac{K_d\,\tau_d\,(\tau_p - \tau_d)}{\tau_d\,(1 + K_p\,K_c) - \tau_p}\,\dfrac{1}{\tau_d s + 1}\right. \\ & & \left. + \dfrac{K_d\,K_p\,K_c\,\tau_p^2}{(\tau_p - \tau_d\,(1 + K_p\,K_c))\,(1 + K_p\,K_c)}\,\dfrac{1}{\tau_p s + 1 + K_p\,K_c}\right) \\ & = & A\left(\dfrac{c_1}{s} + \dfrac{c_2}{\tau_d s + 1} + \dfrac{c_3}{\tau_p s + 1 + K_p\,K_c}\right) \end{array} \tag{2.79}$$

hence the closed-loop response

$$y(t) = A\left[c_1 + c_2\,\exp(-t/\tau_d) + c_3\,\exp(-t\,(1 + K_p\,K_c)/\tau_p)\right] \tag{2.80}$$

In the absence of a controller, the process is stable and the output tends towards AK_d. When the proportional controller is introduced, the process remains stable; for a unit step disturbance, the output tends towards a new limit AK'_d and deviates with respect to the set point from this value AK'_d.
When the gain of the proportional controller increases, the deviation output-set point decreases and the influence of the disturbance decreases too.

Consider now a disturbance impulse variation of amplitude A. The closed-loop response to this disturbance is then equal to

$$Y(s) = A\,\frac{K_d}{\tau_d s + 1}\,\frac{\tau_p s + 1}{\tau_p s + 1 + K_p K_c} \tag{2.81}$$

Application of the final value theorem gives

$$\lim_{t\to\infty} y(t) = \lim_{s\to 0}[s\,Y(s)] = 0 \tag{2.82}$$

thus impulse disturbances are rejected by a simple proportional controller.

Second-Order Systems

Again, the actuator and measurement gains and transfer functions are taken to be equal to 1.

In the case of a second-order system, the process transfer function is

$$G_p(s) = \frac{K_p}{\tau_p^2\, s^2 + 2\,\zeta_p\,\tau_p\, s + 1} \tag{2.83}$$

Assume that the transfer function for the disturbance is also second-order

$$G_d(s) = \frac{K_d}{\tau_d^2\, s^2 + 2\,\zeta_d\,\tau_d\, s + 1} \tag{2.84}$$

The output $Y(s)$ for any set point $Y_r(s)$ and any disturbance $D(s)$ is equal to

$$\begin{aligned} Y(s) \;=\; & \frac{K_p\,K_c}{\tau_p^2\, s^2 + 2\,\zeta_p\,\tau_p\, s + 1 + K_p\,K_c}\,Y_r(s) \\ & + \frac{K_d}{\tau_d^2\, s^2 + 2\,\zeta_d\,\tau_d\, s + 1}\,\frac{\tau_p^2\, s^2 + 2\,\zeta_p\,\tau_p\, s + 1}{\tau_p^2\, s^2 + 2\,\zeta_p\,\tau_p\, s + 1 + K_p\,K_c}\,D(s) \end{aligned} \tag{2.85}$$

Tracking study:
In Eq. (2.85), only the term of the set point variation is concerned. For a set point step variation of amplitude A, $Y_r(s)$ becomes A/s, and $Y(s)$ is decomposed into a sum of two fractions, the natural response of order 2 corresponding to the first factor of the previous expression and the forced response in $1/s$. The closed-loop transfer function remains second-order as in open loop. The period and the damping factor are modified

$$\tau'_p = \frac{\tau_p}{\sqrt{1 + K_p\,K_c}} \qquad \zeta'_p = \frac{\zeta_p}{\sqrt{1 + K_p\,K_c}} \tag{2.86}$$

The steady-state gain becomes

$$K'_p = \frac{K_p\,K_c}{1 + K_p\,K_c} \tag{2.87}$$

Like for the first-order system, a deviation between the set point and the asymptotic response exists (Fig. 2.30), which is all the more important as the gain is low.

Regulation study:

As the influence of disturbance is studied, the second term of Eq. (2.85) is taken into account. The closed-loop response remains second-order as in open loop. The proportional controller is not sufficient to reject the disturbance: a deviation between the set point and the asymptotic value still exists (Fig. 2.32).

A proportional controller does not change the order of the process; the steady-state gain is modified, decreased in two cases (a/ if $K_p > 1$, or b/ if $K_c > 1/(1 - K_p)$ when $K_p < 1$): the time constants also decrease.

2.3.3 Influence of Integral Action

The study is similar to that realized in the case of the proportional controller and will be consequently less detailed.

The transfer function of a PI controller is equal to:

$$G_c(s) = K_c \left(1 + \frac{1}{\tau_I\,s}\right) \tag{2.88}$$

First-Order Process and Influence of Pure Integral Action

Though integral action is never used alone, in this section, in order to characterize its influence, we first assume that the controller is pure integral and has the following transfer function

$$G_c(s) = \frac{K_c}{\tau_I\,s} \tag{2.89}$$

In the case of a first-order process, the response $Y(s)$ to a set point or disturbance variation is equal to

$$Y(s) = \frac{\dfrac{K_p}{1+\tau_p\,s}\,\dfrac{K_c}{\tau_I\,s}}{1 + \dfrac{K_p}{1+\tau_p\,s}\,\dfrac{K_c}{\tau_I\,s}}\,Y_r(s) + \frac{\dfrac{K_d}{1+\tau_d\,s}}{1 + \dfrac{K_p}{1+\tau_p\,s}\,\dfrac{K_c}{\tau_I\,s}}\,D(s) \tag{2.90}$$

or

$$Y(s) = \frac{1}{\dfrac{\tau_p\,\tau_I}{K_p\,K_c}\,s^2 + \dfrac{\tau_I}{K_p\,K_c}\,s + 1}\,Y_r(s) + \frac{K_d}{\tau_d\,s+1}\,\frac{(\tau_p\,s+1)\,\tau_I\,s}{\tau_I\tau_p\,s^2 + \tau_I\,s + K_p\,K_c}D(s) \tag{2.91}$$

The integral controller has modified the system order: the transfer function of the closed-loop system is now of order 2, i.e. larger by one unity than the order of the open-loop system. The natural period of the closed-loop system is equal to

$$\tau'_p = \sqrt{\frac{\tau_p \, \tau_I}{K_p \, K_c}} \tag{2.92}$$

and the damping factor

$$\zeta'_p = \frac{1}{2}\sqrt{\frac{\tau_I}{\tau_p \, K_p \, K_c}} \tag{2.93}$$

As the response of a first-order process in open loop becomes second-order in closed loop, its dynamics is completely different. According to the value of the damping factor ζ', the response will be overdamped, or underdamped, possibly explosive. If the controller gain is increased, keeping constant the integral time constant, the natural period and the damping factor decrease, thus the response will be less sluggish, but the displacement will move progressively from overdamped responses towards oscillatory responses.

It is interesting to study the tracking, i.e. the response to a set point variation. The set point undergoes a step variation of amplitude A

$$Y_r(s) = \frac{A}{s} \tag{2.94}$$

hence

$$Y(s) = \frac{1}{\dfrac{\tau_p \, \tau_I}{K_p \, K_c} s^2 + \dfrac{\tau_I}{K_p \, K_c} s + 1} \, \frac{A}{s} \tag{2.95}$$

To find the asymptotic behaviour, the final value theorem gives

$$\lim_{t \to \infty} y(t) = \lim_{s \to 0} [s \, Y(s)] = A \tag{2.96}$$

thus the limit of $y(t)$ is equal to A, the set point value. We thus find the important result that the integral action eliminates the asymptotic deviation. The value of the set point is reached faster when the gain is high, but at the expense of oscillatory responses. According to the type of controlled variable, it is preferable to rather choose an overdamped response (not going beyond the set point, e.g. for a chemical reactor which could undergo runaway above some safety temperature) or oscillatory (rapidly reach a state close to the set point).

Let us study in a similar way the regulation, thus the influence of a disturbance. Consider a disturbance step change of amplitude A that could not be rejected by the proportional controller

$$D(s) = \frac{A}{s} \tag{2.97}$$

hence

$$Y(s) = \frac{K_d}{\tau_d \, s + 1} \, \frac{(\tau_p \, s + 1) \, \tau_I \, s}{\tau_I \tau_p \, s^2 + \tau_I \, s + K_p \, K_c} \, \frac{A}{s} \tag{2.98}$$

The final value theorem gives

$$\lim_{t\to\infty} y(t) = \lim_{s\to 0}[s\,Y(s)] = 0 \tag{2.99}$$

thus step-like disturbances which were not rejected by a proportional controller are perfectly rejected owing to the integral action.

Of course, impulse disturbances are also rejected by the PI controller.

First-Order Process with PI Controller

The transfer function of the PI controller is equal to

$$G_c(s) = K_c\left(1 + \frac{1}{\tau_I\, s}\right) \tag{2.100}$$

hence the general closed-loop response to set point and disturbance variations

$$Y(s) = \frac{\dfrac{K_p}{1+\tau_p\, s}\, K_c\left(1+\dfrac{1}{\tau_I\, s}\right)}{1+\dfrac{K_p}{1+\tau_p\, s}\, K_c\left(1+\dfrac{1}{\tau_I\, s}\right)}\, Y_r(s) + \frac{\dfrac{K_d}{1+\tau_d\, s}}{1+\dfrac{K_p}{1+\tau_p\, s}\, K_c\left(1+\dfrac{1}{\tau_I\, s}\right)}\, D(s) \tag{2.101}$$

or

$$\begin{aligned} Y(s) \;=\; & \frac{\tau_I\, s+1}{\dfrac{\tau_p\,\tau_I}{K_p\,K_c}s^2 + \tau_I\,\dfrac{K_p\,K_c+1}{K_p\,K_c}s+1}\, Y_r(s) + \\ & \frac{K_d}{\tau_d\, s+1}\,\frac{\dfrac{\tau_I}{K_p\,K_c}\, s\,(1+\tau_p\, s)}{\dfrac{\tau_p\,\tau_I}{K_p\,K_c}s^2 + \tau_I\,\dfrac{K_p\,K_c+1}{K_p\,K_c}s+1}\, D(s) \end{aligned} \tag{2.102}$$

Compared to the proportional action alone, the order of each transfer function output/set point or output/disturbance increases by one unit.

Previously drawn conclusions for the integral action alone remain globally true:

- In tracking, during a set point step variation (Fig. 2.29), the output tends towards the set point even for low controller gain.
- In regulation, impulse disturbances are of course rejected, but also step disturbances (Fig. 2.31).

To be convinced, it suffices to use the final value theorem.

Second-Order Process

Similarly, in the case of a second-order process in open loop, the closed-loop output with a pure integral controller would be of the immediately next order, i.e. a third-order.

The PI controller used in the case of the tracking corresponding to Fig. 2.24 leads to oscillations all the more important for the second-order system in that

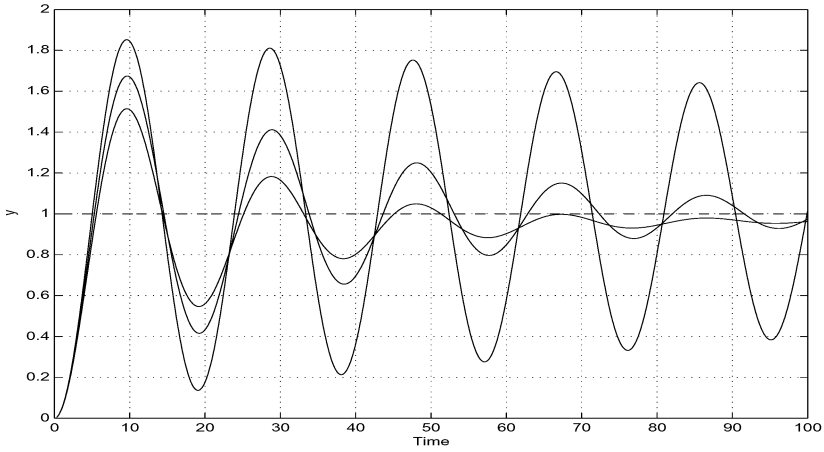


Figure 2.24: Response of a second-order system ($K_p = 5, \tau_p = 10, \zeta_p = 0.5$) to a set point unit step with influence of the integral time constant (PI controller $K_c = 2, \tau_I = 10 \text{ or } 20 \text{ or } 100$). When τ_I increases, the oscillation amplitude decreases

the integral time constant is lower, and thus that the integral gain increases. The deviation with respect to the set point is cancelled by the integral action either for a set point variation (Fig. 2.30), or a disturbance variation (Fig. 2.32). With the overshoot increasing with the integral gain, it will be frequently necessary not to choose too large an integral gain. In those figures, the gain and the integral time constant have not been optimized, as the objective was only to display the influence of the integral action.

2.3.4 Influence of Derivative Action

The transfer function of a pure ideal derivative controller is equal to

$$G_c(s) = K_c \, \tau_D \, s \tag{2.103}$$

This transfer function is improper and the ideal PID controller (Fig. 2.25) of transfer function

$$G_c(s) = K_c \left(1 + \frac{1}{\tau_I s} + \tau_D \, s \right) \tag{2.104}$$

is not realizable, as the numerator degree is larger than the denominator degree because of the ideal derivative action term. If this controller were used as such, it would amplify high-frequency noise because its amplitude ratio is unlimited at high frequency (see frequency analysis, Chap. 5).

The following study simply aims to demonstrate the characteristics of pure derivative action.

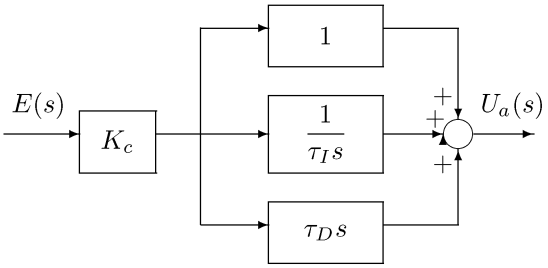


Figure 2.25: Block diagram of the ideal PID controller

First-Order Process and Pure Derivative Action

In the case of a first-order process, if one only looks at the influence of derivative action with the controller given by Eq. (2.103), the response $Y(s)$ to a set point or disturbance variation is equal to

$$\begin{aligned} Y(s) &= \frac{\dfrac{K_p}{\tau_p\, s+1}\, K_c\, \tau_D\, s}{1+\dfrac{K_p}{\tau_p\, s+1}\, K_c\, \tau_D\, s} Y_r(s) + \frac{\dfrac{K_d}{\tau_d\, s+1}}{1+\dfrac{K_p}{\tau_p\, s+1}\, K_c\, \tau_D\, s} D(s) \\ &= \frac{K_p\, K_c\, \tau_D\, s}{(\tau_p + K_p\, K_c\, \tau_D)s+1} Y_r(s) + \frac{K_d}{\tau_d\, s+1}\, \frac{\tau_p\, s+1}{(\tau_p + K_p\, K_c\, \tau_D)s+1} D(s) \end{aligned} \tag{2.105}$$

Transfer functions are first-order as in open loop, thus the derivative action has no influence on the system order. On the other hand, the derivative action introduces a lead term in the numerator. The closed-loop time constant is equal to

$$\tau_p' = \tau_p + K_p\, K_c\, \tau_D \tag{2.106}$$

and thus is increased with respect to the open loop; the closed-loop response will be slower than the open-loop one, and this effect increases with the derivative controller gain. This will help to stabilize the process if the latter shows tendencies to oscillations in the absence of the derivative action.

First-Order Process with Real PID Controller

Compared to the PI controller previously studied, a physically realizable derivative action is introduced by the following real PID controller (Fig. 2.26) of transfer function

$$G_c(s) = K_c \left(\frac{\tau_I s+1}{\tau_I s}\right) \left(\frac{\tau_D s+1}{\beta\, \tau_D s+1}\right) \tag{2.107}$$

which is physically realizable. This transfer function can be seen as the filtering of an ideal PID controller by a first-order filter. In the case of a pneumatic PID controller, β is included between 0.1 and 0.2. For the electronic PID controller, one sets $0 < \beta \ll 1$.

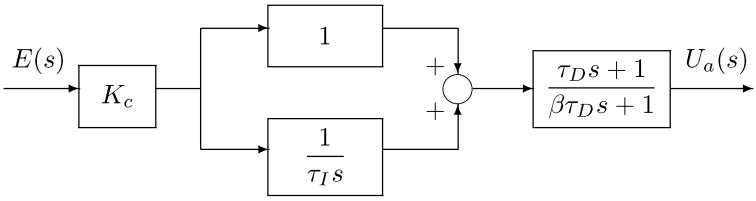


Figure 2.26: Block diagram of the real PID controller given by Eq. (2.107)

A real PID controller (Fig. 2.27) can also respond to the following slightly different equation, which is frequently used

$$G_c(s) = K_c \left(1 + \frac{1}{\tau_I s} + \frac{\tau_D s}{\frac{\tau_D}{N} s + 1} \right) \tag{2.108}$$

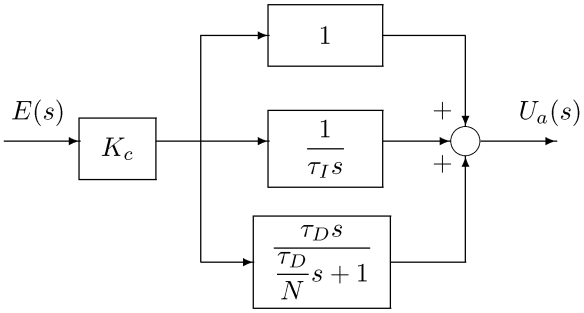


Figure 2.27: Block diagram of the real PID controller given by Eq. (2.108)

In the case of studied first-order processes, the derivative action in the PID controller does not seem to add an important effect with respect to integral action alone, as the studied process already presents a closed-loop overdamped behaviour with the PI controller. If for other parameter values the closed-loop behaviour had been underdamped, the addition of derivative action would have allowed considerable decrease of oscillations which would have become acceptable as in the following case of the second-order process (Fig. 2.28). The influence of the derivative action is clearer in response with respect to a disturbance step variation (Fig. 2.31) than in response with respect to a set point step variation (Fig. 2.29). It is shown that the overshoot is decreased. The derivative action thus brings a stabilizing influence with respect to integral action.

Second-Order Process

In the case of a second-order process and pure integral action, the response $Y(s)$ to a set point variation is equal to

$$Y(s) = \frac{G_p\, K_c\, \tau_D\, s}{1 + G_p\, K_c\, \tau_D\, s} Y_r(s) = \frac{K_p\, K_c\, \tau_D\, s}{\tau^2\, s^2 + 2\,\zeta\,\tau\, s + 1 + K_p\, K_c\, \tau_D\, s} Y_r(s) \quad (2.109)$$

The closed-loop response is second-order as it was in open loop. The derivative controller does not modify the order of the response

$$Y(s) = \frac{K_p\, K_c\, \tau_D\, s}{\tau^2\, s^2 + (2\,\zeta\,\tau\, + K_p\, K_c\, \tau_D) s + 1} Y_r(s) \quad (2.110)$$

In this case, the time constant τ remains the same while the damping factor of the closed-loop response is modified with respect to the open-loop damping factor and becomes

$$\zeta_p' = \frac{2\,\zeta_p\,\tau + K_p\, K_c\, \tau_D}{2\,\tau} \quad (2.111)$$

The response is thus more damped in closed loop than in open loop, and this damping increases with the gain K_c of the derivative controller and with the derivative time constant.

Globally, the same effect is noticed with a real PID controller of transfer function given by Eq. (2.107).

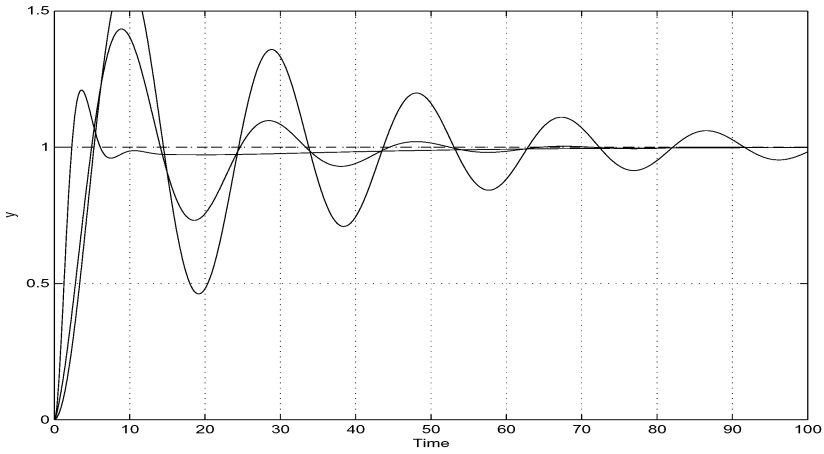


Figure 2.28: Response of a second-order system ($K_p = 5, \tau_p = 10, \zeta_p = 0.5$) to a set point unit step (real PID controller with influence of the derivative time constant τ_D: $K_c = 2, \tau_I = 20, \tau_D = 0.1$ or 1 or 10, $\beta = 0.1$). When τ_D increases, oscillations decrease

Compared to the integral action which cancels asymptotic deviation but leads to strong oscillations, the addition of real derivative action strongly decreases oscillations which become acceptable, all the more so as the derivative time

constant τ_D is higher (Fig. 2.28). However, it must be noted that the increase of derivative action tends to increase measurement noise and that this effect is not wished, so that a too large value of τ_D must be avoided in practice. The derivative action brings a stabilizing effect with respect to integral action. The overshoot is also decreased. These two effects are clear in the study of the influence of either a set point step variation (Fig. 2.30) or a disturbance step variation (Fig. 2.32).

2.3.5 Summary of Controllers Characteristics

A proportional (P) controller contains only one tuning parameter: the controller gain. The asymptotic output presents a deviation from the set point, which can be decreased by increasing the controller gain. The use of too large gains can make the process unstable due to neglected dynamics or time delays.

A proportional-integral (PI) controller presents the advantage of integral action leading to the elimination of the deviation between the asymptotic state and the set point. The response is faster when the gain increases, and can become oscillatory. For large values of the gain, the behaviour may even become unstable. The decreasing of the integral time constant increases the integral gain and makes the response faster. Because of the integral term, the PI controller may present a windup effect if the control variable u becomes saturated. In this case, the integral term becomes preponderant and needs time to be compensated. It is preferable to use an anti-windup system (Sect. 4.6.4).

The proportional-integral-derivative (PID) controller presents the same interest as the PI with respect to the asymptotic state. Furthermore, the derivative action allows a faster response without needing to choose too high gains as for a PI controller. This derivative action thus has a stabilizing effect.

The ideal PID controller is indeed replaced by a real PID controller, the transfer function of which given by Eq. (2.107) or (2.108) is physically realizable. In the case of a pneumatic PID controller, β is included between 0.1 and 0.2. For the PID electronic controller, one sets $0 < \beta \ll 1$.

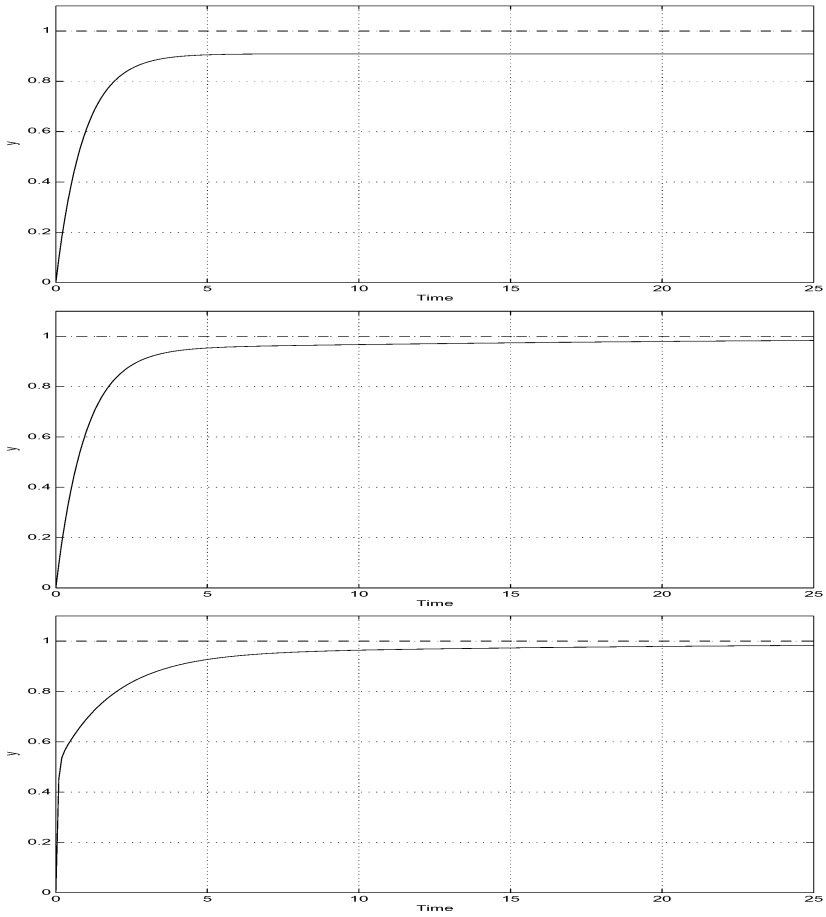


Figure 2.29: Comparison of the influence of the controller type on the response of a first-order system ($K_p = 5, \tau_p = 10$) to a set point unit step. (Proportional: $K_c = 2$ (top). Proportional-integral: $K_c = 2, \tau_I = 20$ (middle). Real proportional-integral-derivative: $K_c = 2, \tau_I = 20, \tau_D = 1, \beta = 0.1$ (bottom))

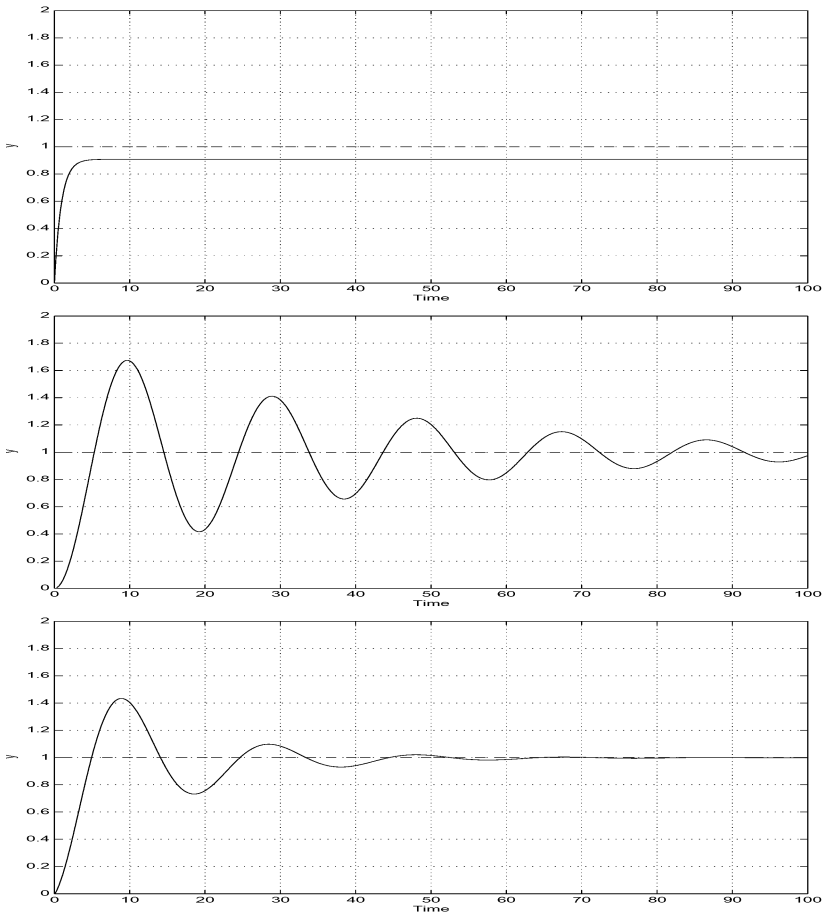


Figure 2.30: Comparison of the influence of the controller type on the response of a second-order system ($K_p = 5, \tau_p = 10, \zeta_p = 0.5$) to a set point unit step. (Proportional: $K_c = 2$ (top). Proportional-integral: $K_c = 2, \tau_I = 20$ (middle). Real proportional-integral-derivative: $K_c = 2, \tau_I = 20, \tau_D = 1, \beta = 0.1$ (bottom))

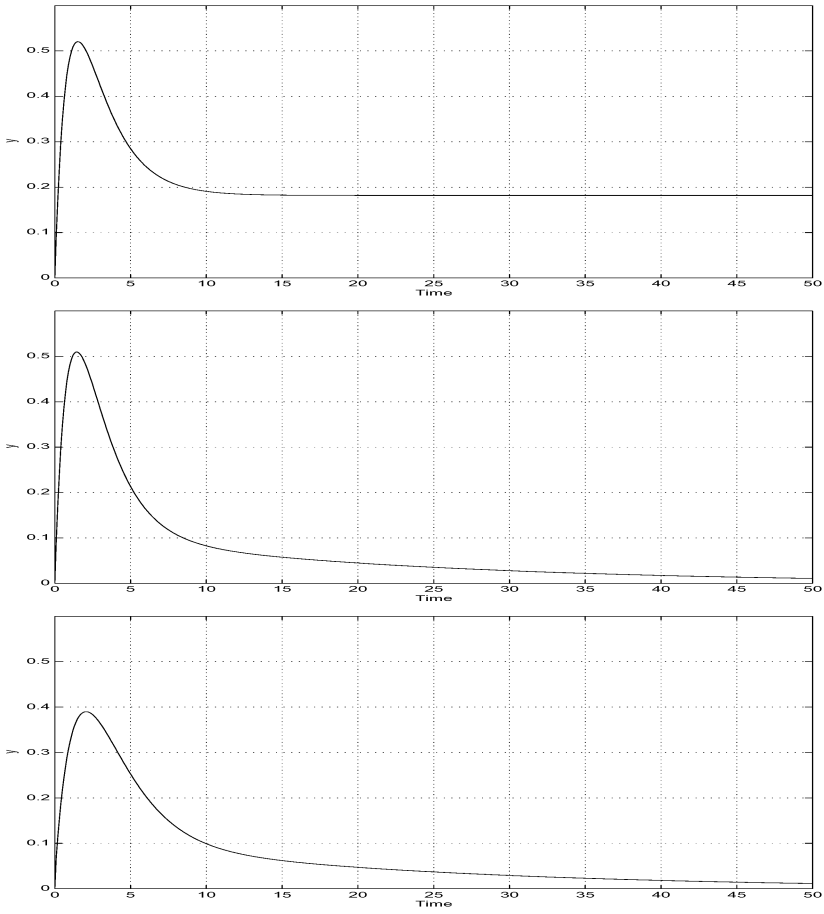


Figure 2.31: Comparison of the influence of the controller type on the response of a first-order system ($K_p = 5, \tau_p = 10 : K_d = 2, \tau_d = 2$) to a disturbance unit step. (Proportional: $K_c = 2$ (top). Proportional-integral: $K_c = 2, \tau_I = 20$ (middle). Real proportional-integral-derivative: $K_c = 2, \tau_I = 20, \tau_D = 1, \beta = 0.1$ (bottom))

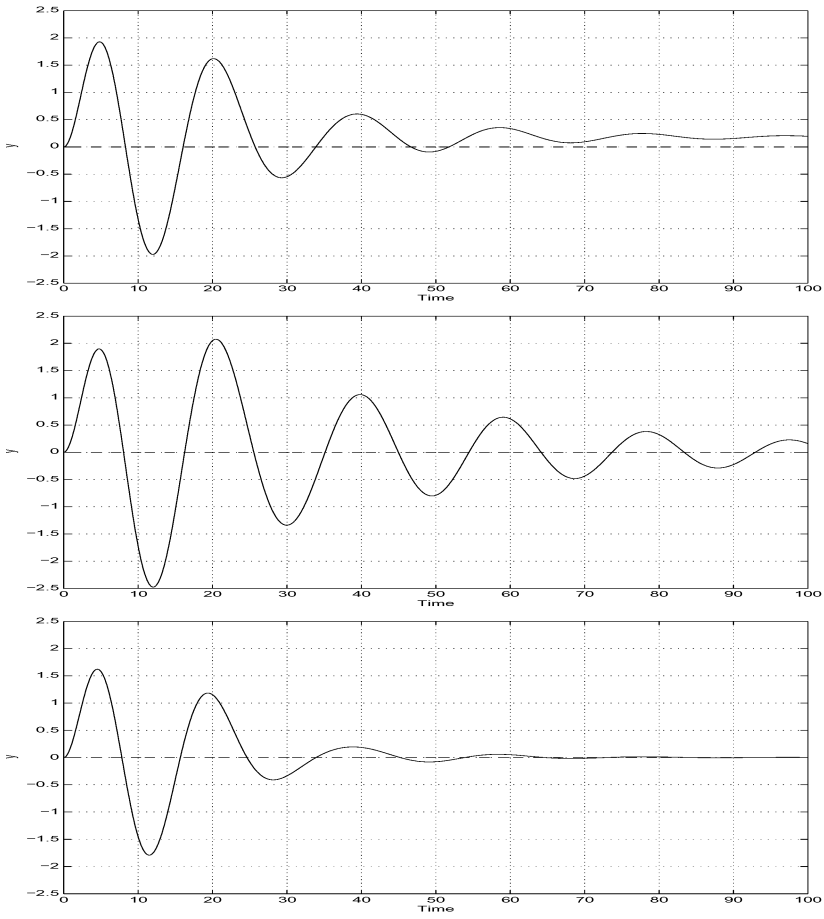


Figure 2.32: Comparison of the influence of the controller type on the response of a second-order system ($K_p = 5, \tau_p = 10, \zeta_p = 0.5 : K_d = 2, \tau_d = 2, \zeta_d = 0.25$) to a disturbance unit step. (Proportional: $K_c = 2$ (top). Proportional-integral: $K_c = 2, \tau_I = 20$ (middle). Real proportional-integral-derivative: $K_c = 2, \tau_I = 20, \tau_D = 1, \beta = 0.1$ (bottom))

References

D.M. Considine, editor. *Process/Industrial Instruments and Controls Handbook*. McGraw-Hill, New York, 5th edition, 1999.

T.E. Marlin. *Process Control. Designing Processes and Control Systems for Dynamic Performance*. McGraw-Hill, Boston, 2000.

N. Midoux. *Mécanique et Rhéologie des Fluides en Génie Chimique*. Lavoisier, Paris, 1985.

D.E. Seborg, T.F. Edgar, and D.A. Mellichamp. *Process Dynamics and Control*. Wiley, New York, 1989.

P. Thomas. *Simulation of Industrial Processes for Control Engineers*. Butterworth-Heinemann, Oxford, 1999.

Chapter 3

Stability Analysis

The stability analysis relies on the same mathematical concepts, whether the system is in open or closed loop. Naturally, functions, variables, matrices, etc., which will be the object of the study, will be different. The cases of linear and nonlinear systems will be studied separately.

A system is stable if a bounded output corresponds to a bounded input in the limits of the problem physics (a curve of undamped sinusoidal allure is not bounded): it is often called BIBO. (Bounded input, bounded output).

However, the previous definition of stability is not totally satisfactory for a closed-loop system: indeed, not only the **external stability** should be ensured, but also the **internal stability**, i.e. all the internal signals inside the closed loop must be bounded (Vidyasagar, 1985). Note that when the used controller is stable, then external stability induces internal stability.

3.1 Case of a System Defined by its Transfer Function

In Sect. 1.9.5, the response of a linear system, defined by its transfer function

$$G(s) = \frac{N(s)}{D(s)} \quad ; \quad \text{degree of } N \leq \text{degree of } D \tag{3.1}$$

to an input step has been studied. This allows us to demonstrate the conditions of stability:

- a system is stable when the poles of its transfer function (roots of the denominator $D(s)$) are all negative real or complex with negative real part. These poles will be said to be stable.
- if a complex pole has its real part equal to zero, the system is called marginally stable.
- a system is unstable when one or several poles of its transfer function are positive real or complex with a positive real part. This type of poles is said to be unstable.

A system is minimum-phase when all zeros of its transfer function (roots of numerator $N(s)$) are all negative real or complex with a negative real part. When there are one or several positive real or complex with positive real part zeros, the system is called nonminimum-phase.

3.2 State-Space Analysis

3.2.1 General Analysis for a Continuous Nonlinear System

Consider a physical system, the time evolution of which is described by a nonlinear ordinary differential equation of order n as

$$\frac{d^n x}{dt^n} = x^{(n)} = f(x, x^{(1)}, \ldots, x^{(n-1)}, t) \tag{3.2}$$

Introducing new variables $x_i = x^{(i-1)}$ $(1 \leq i \leq n)$, the previous differential equation can be written as a set of first-order ordinary differential equations

$$\begin{cases} \dot{x}_1 & = x_2 \\ \vdots & \\ \dot{x}_{n-1} & = x_n \\ \dot{x}_n & = f(x_1, \ldots, x_n, t) \end{cases} \tag{3.3}$$

which is the description of this nonlinear system in state space; variables x_i are the system states.

Frequently, a nonlinear system will be indeed represented by a system of ordinary differential equations of the form

$$\begin{cases} \dot{x}_1 & = f_1(x_1, \ldots, x_n, t) \\ \vdots & \\ \dot{x}_n & = f_n(x_1, \ldots, x_n, t) \end{cases} \tag{3.4}$$

which we will denote by

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}) \quad , \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{3.5}$$

Solutions $\phi_t(\boldsymbol{x}_0)$ of the previous differential system are called integral curves of the vector field $\boldsymbol{f}$ and depend on the initial states $\boldsymbol{x}_0$. These curves can be parameterized with respect to t and are also called trajectories; they represent the locus of the successive states of the system. They verify $\phi_t(\boldsymbol{x}_0, t_0) = \boldsymbol{x}_0$.

When the time variable t does not intervene explicitly, the system is called autonomous. Any autonomous nonlinear continuous system can be represented by the following set of ordinary differential equations

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) \quad , \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{3.6}$$

A point $\boldsymbol{x}^*$ is called stationary (or singular) (or equilibrium) if

$$\boldsymbol{f}(\boldsymbol{x}^*) = 0 \tag{3.7}$$

If the stationary point is not placed at the origin, it can be brought to the origin by making the translation $\boldsymbol{x}' = \boldsymbol{x} - \boldsymbol{x}^*$. Thus, discussions and theorems consider the origin as the stationary point around which the stability is studied.

It is possible to define rigorously from a mathematical point of view several kinds of stability:

Lyapunov Stability :

The stationary point at the system origin (3.5) is stable if for all $\epsilon > 0$, there exists a scalar $\alpha(\epsilon, t_0)$ such that $\|\boldsymbol{x}_0\| < \alpha$ implies $\lim \|\phi_t(\boldsymbol{x_0}, t)\| < \epsilon$, for all $t > t_0$.

Lyapunov stability means that all trajectories remain in a neighbourhood of the origin. The considered norm is the Euclidean norm[1].

Asymptotic stability:

The stationary point at the system origin (3.5) is quasi-asymptotically stable if there exists a scalar β such that $\|\boldsymbol{x}_0\| < \beta$ induces $\lim_{t\to\infty} \|\phi_t(\boldsymbol{x_0}, t)\| = 0$.

The asymptotic stability furthermore implies that the origin must also be stable, as all trajectories tend towards the origin.

Instability means that trajectories move away from origin.

First-Order Stability

The continuous-time nonlinear system is linearized around the stationary point $\boldsymbol{x}^*$, thus its linear approximation

$$\delta\dot{\boldsymbol{x}} = \boldsymbol{A}\delta\boldsymbol{x} \qquad \text{with:} \quad \delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}^* \tag{3.8}$$

governs the time evolution when the system is perturbed from $\delta\boldsymbol{x}$ around the stationary point $\boldsymbol{x}^*$. $\boldsymbol{A}$ is the Jacobian matrix or stability matrix

$$\boldsymbol{A} = \left[\begin{array}{ccc} \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_1} & \cdots & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_n} \end{array} \right]_{\boldsymbol{x}^*} \tag{3.9}$$

If the eigenvalues[2] λ_i of $\boldsymbol{A}$ are distinct, if the associated eigenvectors are noted η_i, the trajectory can be described (Parker and Chua, 1989) at first-order by

$$\phi_t(\boldsymbol{x}^* + \delta\boldsymbol{x}) = \boldsymbol{x}^* + c_1 \exp(\lambda_1 t)\eta_1 + \ldots + c_n \exp(\lambda_n t)\eta_n \tag{3.10}$$

where c_i are constant coefficients satisfying the initial condition. A stationary point such that no eigenvalue has a real part equal to zero is called a hyperbolic

[1] The Euclidean norm of a vector is the square root of the sum of the squares of its elements, i.e. the "length" of a vector.

[2] The eigenvalues λ_i of a matrix $\boldsymbol{A}$ are the roots of its characteristic polynomial of degree n, equal to the determinant of the matrix $\boldsymbol{A} - \lambda\boldsymbol{I}$.

point. Only hyperbolic points are considered in this framework (Guckenheimer and Holmes, 1986). Two cases are possible:

- If the eigenvalue λ_i is real, it represents a contraction velocity if $\lambda_i < 0$ and an expansion one if $\lambda_i > 0$, in the neighbourhood of the stationary point $\boldsymbol{x}^*$ in direction η_i,
- If the eigenvalue λ_i is complex, its conjugate complex $\bar{\lambda}_i$ is also an eigenvalue as $\boldsymbol{A}$ is a real matrix; the conjugate vector $\bar{\eta}_i$ is the eigenvector associated with the eigenvalue $\bar{\lambda}_i$. The contribution

$$c_i \exp(\lambda_i t)\eta_i + \bar{c}_i \exp(\bar{\lambda}_i t)\bar{\eta}_i \tag{3.11}$$

  describes a spiral in the plane generated by $\mathcal{R}e(\eta_i)$ and $\mathcal{I}m(\eta_i)$. The real part of eigenvalue λ_i represents a contraction velocity of the spiral if $\mathcal{R}e(\lambda_i) < 0$ and an expansion one if $\mathcal{R}e(\lambda_i) > 0$.
- If all eigenvalues of $\boldsymbol{A}$ have their real part strictly negative, the stationary point is asymptotically stable.
- When all eigenvalues have their real part positive, the stationary point is unstable.
- When some eigenvalues have their real part negative and others their real part positive, the stationary point is a saddle point.
- When all eigenvalues are real, the stationary point is a node.
- When all eigenvalues are in the same complex half-plane (either left or right), and some of them are complex, the stationary point is a focus.

In dimension 2, when the eigenvalues are equal and when the matrix $\boldsymbol{A}$ is diagonalizable, the stationary point is a focus ; if matrix $\boldsymbol{A}$ is not diagonalizable, the stationary point is a node (d'Andrea Novel and de Lara, 1994).

In the case of a stable node, this stationary point is an attractor: if a trajectory originates from a point situated in the neighbourhood of this point, it converges directly towards the attractor without turning around it.

In the case of a stable focus, this stationary point is an attractor: if a trajectory originates from a point situated in the neighbourhood of this point, it converges towards the attractor, turning around it.

In the case of an unstable node and of an unstable focus, trajectories move away from the stationary point.

Stability properties are more easily visualized in a plane ($n = 2$) which easily allows a graphical representation. This representation made in the phase plane or phase space is called a phase portrait and the variables x_i are the phase space coordinates. When $n > 2$, the followed space directions interfere and strongly complicate the discussion (Gilmore, 1981).

Asymptotic Stability Domain

The asymptotic stability domain (Pellegrini et al., 1988) of a point $\boldsymbol{x}^*$ is the set of regular points $\boldsymbol{x}_0$ such that

$$\lim_{t \to \infty} \boldsymbol{x}(\boldsymbol{t}, \boldsymbol{x_0}) = \boldsymbol{x}^* \tag{3.12}$$

The frontier of an asymptotic stability domain is formed by the set of trajectories. If the asymptotic stability domain is limited, its frontier is formed either by a limit cycle, or by a phase polygon, or by a closed curve of critical points. The frontier can be determined by the following method (Pellegrini et al., 1988):

- The stationary points of the system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ are determined by solving $\boldsymbol{f}(\boldsymbol{x}) = 0$.
- Their stability analysis is realized by calculating the eigenvalues of the linearized system.
- An arbitrarily small stability neighbourhood is determined around each stationary point.
- The system

$$\dot{\boldsymbol{x}} = -\boldsymbol{f}(\boldsymbol{x}) \tag{3.13}$$

  is integrated, starting from the points in the neighbourhood of the stationary points.
- The integration of both concerned systems (3.6) and (3.13) is realized in the neighbourhood of the other critical points.
- Lastly, the asymptotic stability domain is determined by topological considerations.

It must be noted that the integration of system (3.13) is equivalent to a backward integration (with respect to time) of system (3.6). The trajectories of system (3.13) are the same as those of system (3.6), but directed in the opposite way. Moreover, the asymptotically stable stationary points of Eq. (3.6) become unstable for Eq. (3.13).

A limit cycle is stable when close trajectories come near it asymptotically; it is unstable when they move away.

Lyapunov Direct Method

The Lyapunov direct method or Lyapunov second method is an extension of the idea of the mechanical Lagrange energy to study the behaviour of the system with respect to stability without needing to solve its equations (Storey, 1979). In chemical engineering and, in particular, for chemical reactors considered in open or closed loop, some important references should be cited (Aris and Amundson, 1958; Berger and Perlmutter, 1964; Himmelblau and Bischoff, 1968). For linear systems, the construction of a Lyapunov function is easy to realize, seldom for nonlinear systems, but its theoretical interest is important. For a nonlinear system, it is sometimes possible to realize the study in the neighbourhood of an equilibrium point by first making a linearization and then studying the linear approximation.

The Lyapunov function (Fig. 3.1) $V(\boldsymbol{x})$ is a function such that the surfaces defined by $V = \textit{constant}$ are closed, concentric and decrease towards the origin (or the stationary point) when $\|\boldsymbol{x}\| \to 0$. Moreover, the trajectories cross these surfaces towards the inside when the origin is a stable point.

Stability and asymptotic stability theorems:

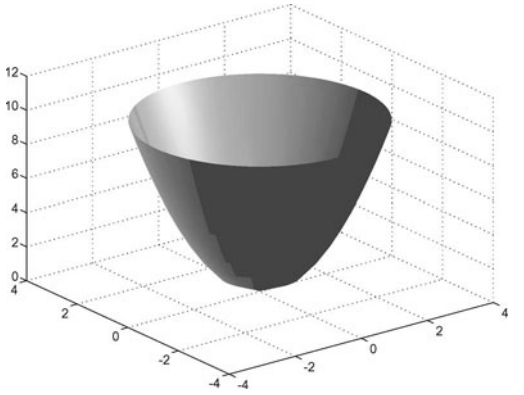


Figure 3.1: Typical shape of a Lyapunov function

- If $V(\boldsymbol{x})$ is definite[3] in the neighbourhood of the origin,
- If $V(\mathbf{0}) = 0$,
- If $\dot{V}(\boldsymbol{x})$ is semidefinite (resp. definite) and of opposite sign to $V(\boldsymbol{x})$,

then the origin is in a stable state (resp. asymptotically stable) of equilibrium of the system.

The function $V(\boldsymbol{x})$ responding to these conditions is a Lyapunov function of the system. Notice that the derivative of $V(\boldsymbol{x}, t)$ is equal to

$$\dot{V}(\boldsymbol{x}, t) = \sum_{i=1}^{n} \frac{\partial V}{\partial x_i} \frac{dx_i}{dt} + \frac{\partial V}{\partial t} \tag{3.14}$$

Asymptotic stability theorem in all space:
the origin is asymptotically stable in all space (for all trajectories coming from all the phase space points) or completely stable if there exists a function $V(\boldsymbol{x})$ such that:

- $V(\boldsymbol{x})$ is positive definite $\forall \boldsymbol{x}$.
- Its derivative $\dot{V}(\boldsymbol{x})$ is negative definite $\forall \boldsymbol{x}$.
- $V(\boldsymbol{x}) \to \infty$ when $\|\boldsymbol{x}\| \to \infty$.

As has already been mentioned, the construction of Lyapunov functions is often delicate, sometimes impossible (Himmelblau and Bischoff, 1968; Storey, 1979; Warden et al., 1964). The linear case is well adapted to this construction. Consider the linear system

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} \tag{3.15}$$

and the quadratic form

$$V(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{B}\boldsymbol{x} \tag{3.16}$$

[3] A function $f(\boldsymbol{x})$ is positive definite in a domain $\mathcal{D}$ around the origin if f and its partial derivatives $\partial f / \partial x_i$ exist and are continuous in $\mathcal{D}$; if furthermore $f(\mathbf{0}) = 0$, $V(\boldsymbol{x}) > 0$ for $x \neq 0$. It is semipositive definite if $V(\boldsymbol{x}) \geq 0$ for $x \neq 0$.

its derivative is equal to

$$\dot{V}(\boldsymbol{x}) = \dot{\boldsymbol{x}}^T \boldsymbol{B}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{B}\dot{\boldsymbol{x}} = \boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{B} + \boldsymbol{B}\boldsymbol{A})\boldsymbol{x} \tag{3.17}$$

Theorem:
The origin is asymptotically stable (the eigenvalues of $\boldsymbol{A}$ have their real part negative) if and only if there exists a matrix $\boldsymbol{B}$ symmetrical positive definite, solution of the Lyapunov equation

$$\boldsymbol{A}^T\boldsymbol{B} + \boldsymbol{B}\boldsymbol{A} + \boldsymbol{C} = 0 \tag{3.18}$$

for all $\boldsymbol{C}$ a positive definite symmetrical matrix.

By using this theorem, the derivative of $V(\boldsymbol{x})$ is equal to

$$\dot{V}(\boldsymbol{x}) = -\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} \tag{3.19}$$

thus is negative definite. It results that the function $V(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}$ satisfying these conditions is a Lyapunov function.

3.2.2 Case of a Linear Continuous System

Example 3.1: **Stability Study of an Harmonic Oscillator**

A remarkable example of stability study allowing visualization is the harmonic oscillator modelled by the second-order ordinary differential equation

$$\tau^2 \frac{d^2x}{dt^2} + 2\,\zeta\,\tau \frac{dx}{dt} + x = 1 \tag{3.20}$$

The stationary point corresponds to $x = 1$. Using this condition, set $x_1 = x - 1$ and $x_2 = \dot{x_1}$. The equation can be written in the phase space as

$$\begin{cases} \dot{x_1} & = x_2 \\ \dot{x_2} & = -(2\,\zeta\,\tau x_2 + x_1)/\tau^2 \end{cases} \tag{3.21}$$

which corresponds to a stable system. Three cases have been studied:

- Underdamped second-order ($\tau = 1,\ \zeta = 0.5$).
- Critically damped second-order ($\tau = 1,\ \zeta = 1$).
- Overdamped second-order ($\tau = 1,\ \zeta = 2$).

with the same initial conditions ($x_1 = 2,\ x_2 = 1$).

For each case, the states have been represented versus time (left of Fig. 3.2) and the phase portrait corresponding to the curves of state x_2 with respect to x_1 (right of Fig. 3.2). In the left-hand figures, it appears that the states converge towards 0 when time becomes large. In the phase portrait, the trajectory tends towards the origin; if the time parameter were represented on the trajectories, the time arrow would be directed towards the origin, i.e. the convergence point. For the underdamped second-order system, the origin is a stable focus. For critically damped and overdamped second-order systems, the origin is a stable node.

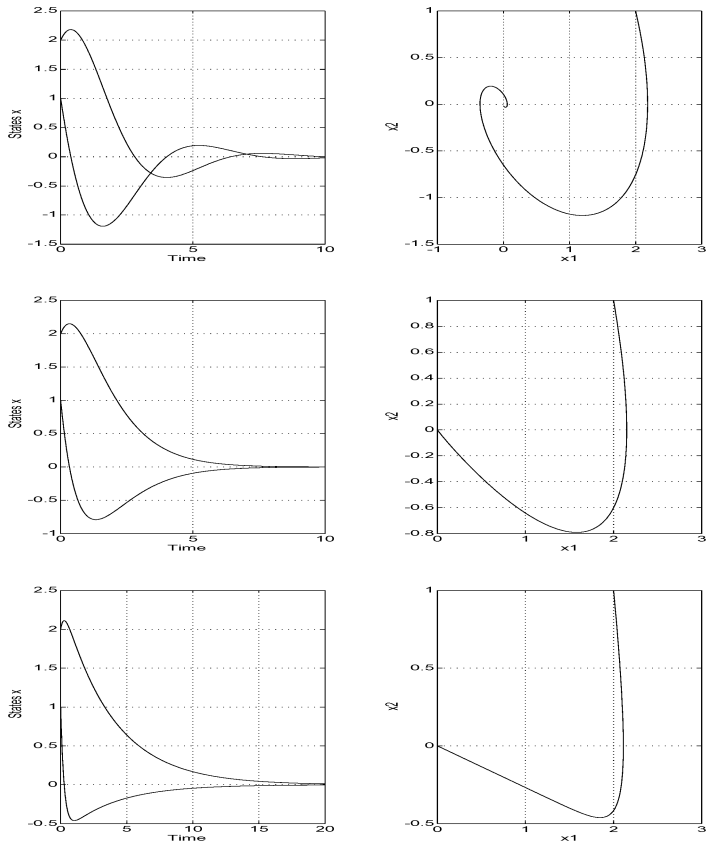


Figure 3.2: Case of a second-order stable linear system. Left-hand figures: representation of the states versus time. Right-hand figures: phase portrait. (Top: $\tau = 1,\ \zeta = 0.5$, middle: $\tau = 1,\ \zeta = 1$, bottom: $\tau = 1,\ \zeta = 2$)

It is possible to study a second-order unstable linear system represented by the following ordinary differential equation

$$\tau^2 \frac{d^2x}{dt^2} + 2\,\zeta\,\tau \frac{dx}{dt} - x = 1 \tag{3.22}$$

This system has no stable stationary point. By performing the same variable change as previously, the system becomes in the state space

$$\left\{ \begin{array}{ll} \dot{x_1} & = -x_2 \\ \dot{x_2} & = -(2\,\zeta\,\tau x_2 + x_1)/\tau^2 \end{array} \right. \tag{3.23}$$

The origin is a stationary point. This system has been studied for four different initial conditions (cases (a), (b), (c), (d) of Fig. 3.3) corresponding to:

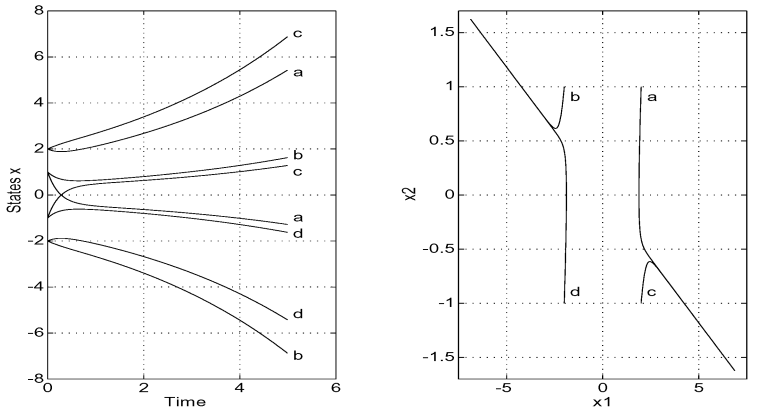


Figure 3.3: Case of a second-order unstable linear system. Left-hand figures: representation of the states versus time. Right-hand figures: phase portrait

(a) $x_1(0) = 2,\ x_2(0) = 1$; (b) $x_1(0) = -2,\ x_2(0) = 1$; (c) $x_1(0) = 2,\ x_2(0) = -1$; (d) $x_1(0) = -2,\ x_2(0) = -1$.

It is not necessary for the time to become important for the states to increase rapidly (in absolute value). In the phase portrait, the points on the trajectories move away from the initial point, with a tendency to come near to the origin and then move away without tending towards any limit. It even appears that asymptotes depend on the initial point. The origin in fact here is a saddle point.

3.2.3 Case of a Nonlinear Continuous System: The Polymerization Reactor

Polymerization reactors are well known for their behaviour, which presents several stationary points because of the high exothermicity of the polymerization reaction.

Example 3.2: **Stability Study of a Polymerization Reactor**

The studied example drawn from Hidalgo and Brosilow (1990) concerns styrene polymerization (Fig. 3.4). The continuous reactor is fed by three independent streams of monomer, initiator and solvent. It is cooled by a heating-cooling fluid circulating in a jacket.

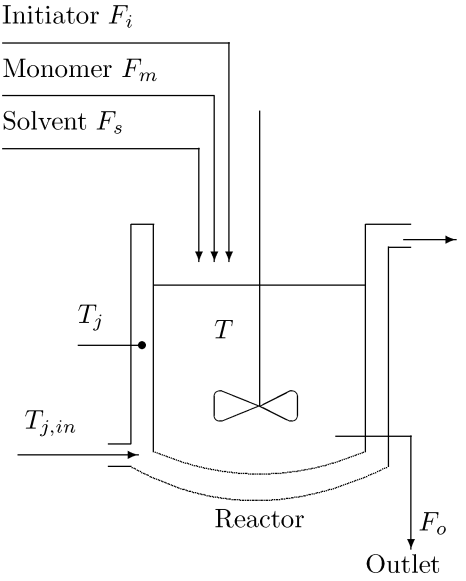


Figure 3.4: Continuous polymerization reactor

The dynamic simulation model reads

$$\begin{cases} \dot{x_1} &= \dfrac{(F_i\, C_{ia} - F_o\, x_1)}{V} - k_d\, x_1 \\ \dot{x_2} &= \dfrac{(F_m\, C_{ma} - F_o\, x_2)}{V} - k_p\, x_2\, \mathcal{R} \\ \dot{x_3} &= \dfrac{F_o\,(T_a - x_3)}{V} - \dfrac{\Delta H}{\rho\, C_p}\, k_p\, x_2\, \mathcal{R} - \dfrac{U\,A}{\rho\, C_p\, V}\,(x_3 - x_4) \\ \dot{x_4} &= \dfrac{F_j (T_{j,in} - x_4)}{V_j} + \dfrac{U\,A}{\rho_j\, C_{pj}\, V_j}\,(x_3 - x_4) \end{cases} \tag{3.24}$$

where the states are, respectively:
x_1: initiator concentration,
x_2: monomer concentration,
x_3: temperature,
x_4: jacket temperature.

The signification of the parameters and variables and their values are given in Table 3.1.

The total outlet volume flow rate F_o is equal to the sum of the inlet flow rates $F_o = F_i + F_m + F_s$. The initiator, the monomer and the solvent are assumed to enter at the same temperature T_a.
The chain concentration of growing polymer is equal to

$$\mathcal{R} = (2\, f\, k_d\, x_1 / k_t)^{0.5} \tag{3.25}$$

The dissociation, propagation and termination rate constants follow the Arrhenius law: $k_i = k_{i0}\, \exp(-E_i/RT)$, ($i = d$, p or t).

Table 3.1: Nominal variables and main parameters of the continuous polymerization reactor

Feed solvent flow rate	$F_s = 0.1275$ l/s
Feed monomer flow rate	$F_m = 0.105$ l/s
Feed initiator flow rate	$F_i = 0.03$ l/s
Feed monomer concentration	$C_{ma} = 8.698$ mol/l
Feed initiator concentration	$C_{ia} = 0.5888$ mol/l
Feed temperature	$T_a = 330$ K
Reactor volume	$V = 3000$ l
Jacket volume	$V_j = 3312.4$ l
Inlet jacket temperature	$T_{j,in} = 295$ K
Density of the reacting mixture × heat capacity	$\rho C_p = 360$
Heat-conducting fluid flow rate	$F_j = 0.131$ l/s
Density of the heat-conducting fluid × heat capacity	$\rho_j C_{pj} = 966.3$
Global heat transfer coefficient × heat exchange area between the jacket and the reactor contents	$UA = 70$
Initiator efficiency	$f = 0.6$
Heat of reaction	$\Delta H = -16700$
Preexponential factor for dissociation	$k_{d0} = 5.95 \times 10^{13}$
Activation energy for dissociation	$E_d/R = 14897$ K
Preexponential factor for propagation	$k_{p0} = 1.06 \times 10^7$
Activation energy for propagation	$E_p/R = 3557$ K
Preexponential factor for termination	$k_{t0} = 1.25 \times 10^9$
Activation energy for termination	$E_t/R = 843$ K

The conversion rate corresponding to the reaction advancement is equal to

$$\chi = (F_m C_{ma} - F_o x_2)/(F_m C_{ma}) \tag{3.26}$$

To demonstrate the multiple stationary states, a series of simulations was performed with the same initial states $x_1 = 0.0486$ mol/l and $x_4 = 316.2$ K, but with a variable monomer concentration x_2 as well as the reactor temperature x_3, which varied from 305 K to 350 K to obtain Fig. 3.5. Fixed initial values were chosen in accordance with the stationary values indicated by Hidalgo and Brosilow (1990). The curves drawn on the figure are the trajectories followed by an operating point with respect to time. One notices that at temperatures lower than or equal to 345 K, trajectories end at the stable stationary point at low conversion $\chi = 0.045$ and low temperature $x_3 = T_r = 323.6$ K. On the other hand, at temperatures higher than or equal to 346 K, trajectories end at the stable stationary point at high conversion $\chi = 0.800$ and high temperature $x_3 = T_r = 406.0$ K. The unstable stationary point thus corresponds to an intermediate temperature around 345 or 346 K. Trajectories tend to come closer to the unstable point that can be graphically approximately situated at $T_r = 344.6$ K, $\chi = 0.24$.

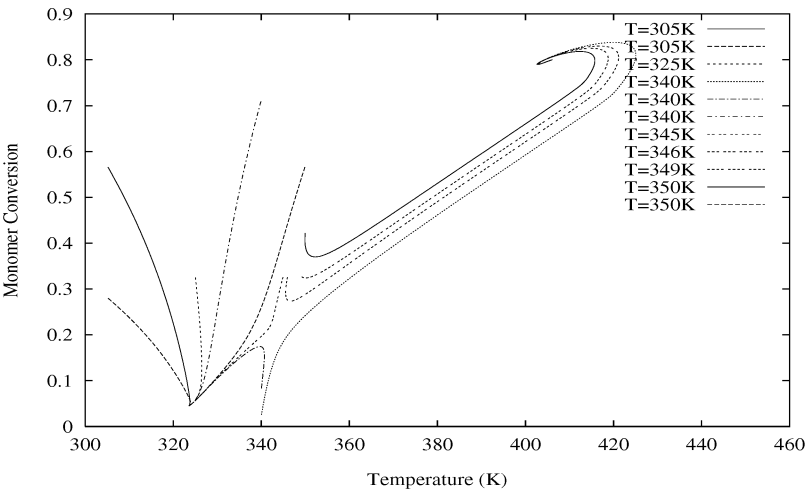


Figure 3.5: Phase portrait: conversion versus temperature for a styrene polymerization reactor

To get the stationary points, it suffices to transform the previous set of ordinary differential equations by cancelling the time derivatives. One thus obtains a nonlinear system of algebraic equations which can be solved by the Newton-Raphson method. The obtained solution depends on the initial point, which was chosen to be equal to $x_1 = 0.05$ mol/l, $x_2 = 2.0$ mol/l and $x_4 = 300$ K, except for the reactor temperature x_3, which varied from 320 K to 400 K. Thus, the two stable stationary points were found:
$x_1 = 0.06683$ mol/l, $x_2 = 3.325$ mol/l, $x_3 = 323.6$ K, $x_4 = 305.2$ K, $\chi = 0.04446$;
$x_1 = 0.0008258$ mol/l, $x_2 = 0.6927$ mol/l, $x_3 = 406.2$ K, $x_4 = 334.6$ K, $\chi = 0.8009$;
as well as the unstable stationary point:
$x_1 = 0.06148$ mol/l, $x_2 = 2.703$ mol/l, $x_3 = 343.1$ K, $x_4 = 312.1$ K, $\chi = 0.2232$.
The two stable points had been previously determined with the help of trajectories: the first solution corresponds to the low conversion point and the second solution to the high conversion point. Moreover, this method allows us to place with accuracy the unstable stationary point.

To determine the frontier separating the domains of asymptotic stability, the technique described in Sect. 3.2.1 was used. Stationary points have just been determined. Starting from the neighbourhood of the unstable saddle point, the integration of the system of the form (3.13), when the polymerization reactor model is symbolized by the form (3.6), was realized. The frontier separating the two domains of asymptotic stability was thus obtained and was thus materialized by the two curves added to Fig. 3.6 compared to Fig. 3.5.

The eigenvalues associated with the linearized system around the stationary point, obtained from (3.24), have been calculated for the three stationary points

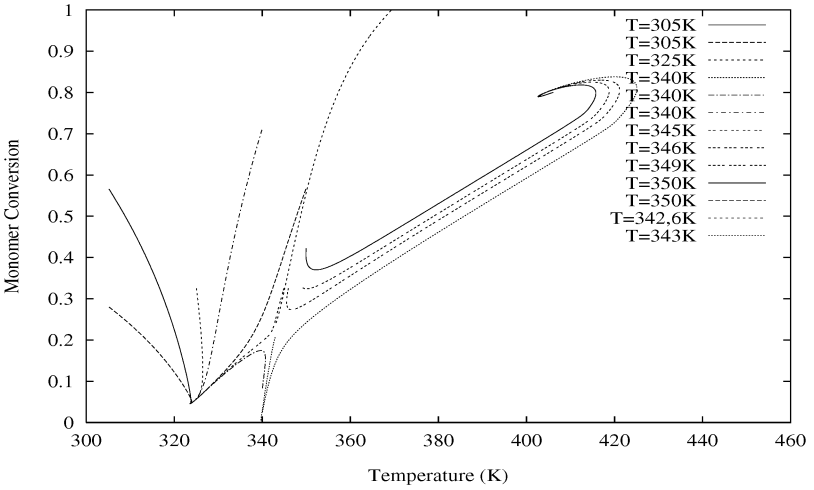


Figure 3.6: Phase portrait: conversion versus temperature for a styrene polymerization reactor. On this figure, to be compared with the previous figure, the frontier between the two domains of asymptotic stability is drawn

and have given the following results:

- Low conversion point:
$$\lambda_1 = -0.366 \times 10^{-4};\ \lambda_2 = -0.88 \times 10^{-4}$$
$$\lambda_3 = -0.1024 \times 10^{-3} + i\,0.63 \times 10^{-5};\ \lambda_4 = -0.1024 \times 10^{-3} - i\,0.63 \times 10^{-5}$$
thus this point is a stable focus.
- High conversion point:
$$\lambda_1 = -0.66 \times 10^{-2};\ \lambda_2 = -0.44 \times 10^{-4}$$
$$\lambda_3 = -0.20 \times 10^{-2} - i\,0.12 \times 10^{-2};\ \lambda_4 = -0.20 \times 10^{-2} + i\,0.12 \times 10^{-2}$$
thus this point is a stable focus.
- Unstable intermediate point:
$$\lambda_1 = 0.99 \times 10^{-4};\ \lambda_2 = -0.945$$
$$\lambda_3 = -0.72 \times 10^{-4} - i\,0.19 \times 10^{-4};\ \lambda_4 = -0.72 \times 10^{-4} + i\,0.19 \times 10^{-4}$$
thus this point is a saddle point.

A classical explanation of the existence of multiple stationary points for the chemical reactor in which an exothermic reaction occurs (case of polymerization reactions) is the following (Van Heerden, 1953). A high temperature induces a high reaction rate. The shape of the curve of heat $\dot{Q}_g$ generated by the chemical reaction with respect to reactor temperature is sigmoidal (Fig. 3.7). In the case of the polymerization reaction studied in this section, the heat $\dot{Q}_g$ generated by the reaction (counted positively as received by the system) is equal to

$$\dot{Q}_g = V\,(-\Delta H\,k_p\,[M]\,\mathcal{P}) \tag{3.27}$$

The reactor is cooled by a cooling fluid circulating in a jacket at mean temperature T_j, and the transferred heat (given to the cooling fluid) through the

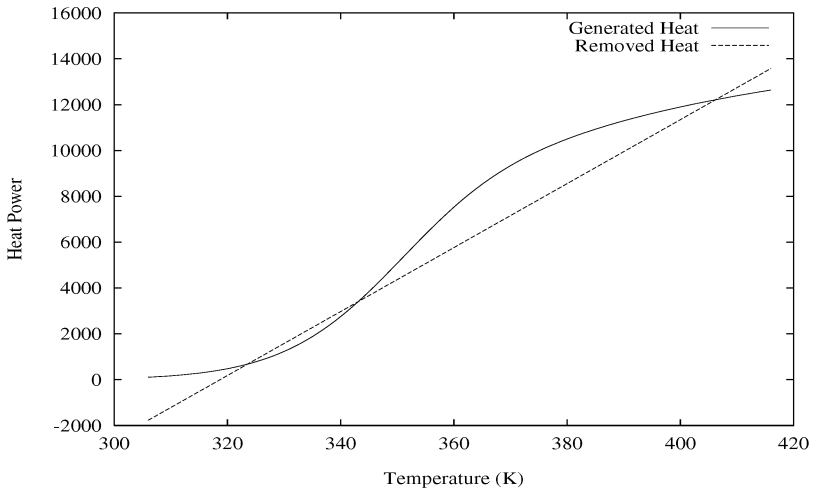


Figure 3.7: Diagram of heat power versus reactor temperature

wall separating the jacket from the reactor is

$$\dot{Q}_{tr} = U\,A\,(T_r - T_j) \tag{3.28}$$

where U is the global heat transfer coefficient through the exchange surface of area A. As the reactor is fed by a stream at temperature T_a lower than the reactor temperature T_r, the cooling capacity of the reactor (counted positively if $T_r > T_j$) is equal to

$$\dot{Q}_r = UA\,(T_r - T_j) + \rho\,C_p\,F_o\,(T_r - T_a) \tag{3.29}$$

with monomer concentration $[M] = x_2$, reactor temperature $T_r = x_3$ and mean temperature of the cooling fluid in the jacket $T_j = x_4$, which gives a straight line in the diagram $(T_r, \dot{Q})$. If the slope of the cooling line is insufficient, the cooling line and the curve of the generated heat present three crossing points which have been previously commented on. The limit corresponds to the tangent at the inflection point of the sigmoid. The slope of the cooling line is a linear function of the inlet temperature T_{ce} of the cooling fluid; it also depends on the flow rate F_c of the cooling fluid and on the global heat transfer coefficient UA.

To obtain point by point the curves of Fig. 3.7, we assumed that the reactor temperature T_r is perfectly controlled, thus T_r is considered no more as a state variable, and that the continuous reactor has reached its steady state. Then it was sufficient to make temperature T_r vary in an adequate domain.

In chemical engineering practice, it is frequently desirable to control the reactor at the unstable stationary point.

3.2.4 State-Space Analysis of a Linear System

Consider a general linear system defined in state space by

$$\left\{ \begin{array}{ll} \dot{\boldsymbol{x}} & = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{B}\,\boldsymbol{u} \\ \boldsymbol{y} & = \boldsymbol{C}\,\boldsymbol{x} + \boldsymbol{D}\,\boldsymbol{u} \end{array} \right. \tag{3.30}$$

and having n_u inputs u_i, n states x_i and n_y outputs y_i. This system thus includes n differential equations, the stability of which must be studied. Indeed, if the states are bounded, the output, which is a linear combination of the states and the inputs, will be bounded.

On the other hand, it is sufficient to study the stability with respect to only one input, as for this linear system, the obtained properties will be able to be extended to all other inputs. Thus, the following system will be studied

$$\left\{ \begin{array}{ll} \dot{x_1} & = a_{11}\,x_1 + \ldots + a_{1n}\,x_n + b_{11}\,u_1 \\ \vdots & \\ \dot{x_n} & = a_{n1}\,x_1 + \ldots + a_{nn}\,x_n + b_{n1}\,u_1 \end{array} \right. \tag{3.31}$$

The input u can be assumed to be zero. Indeed, a constant input corresponding to an input step of amplitude k produces a response which can be decomposed into two parts: the forced response linked to the amplitude of the input and the natural response linked to the system itself. As the problem is to characterize the system, the natural response is sufficient. Under these conditions, exponential solutions of the form $x_i(t) = f_i\,\exp(st)$ of this system are searched for. The system can then be written

$$\left\{ \begin{array}{ll} s\,f_1\,\exp(st) & = a_{11}\,f_1\,\exp(st) + \ldots + a_{1n}\,f_n\,\exp(st) \\ \vdots & \\ s\,f_n\,\exp(st) & = a_{n1}\,f_1\,\exp(st) + \ldots + a_{nn}\,f_n\,\exp(st) \end{array} \right. \tag{3.32}$$

or, after arrangement,

$$[\boldsymbol{A} - s\,\boldsymbol{I}]\,\boldsymbol{f} = 0 \tag{3.33}$$

where $\boldsymbol{f}$ is the vector of coefficients f_i and $\boldsymbol{I}$ is the identity matrix. So that the solution of this system is different from the trivial solution (all s_i zeros), the determinant of $(\boldsymbol{A} - s\,\boldsymbol{I})$ must be zero

$$\det([\boldsymbol{A} - s\,\boldsymbol{I}]) = 0 \tag{3.34}$$

This equation is, in fact, the characteristic polynomial of matrix $\boldsymbol{A}$ and the roots s_i are the eigenvalues of $\boldsymbol{A}$. The state-space stability condition is deduced:
so that a linear system, defined in state space, is stable, it is necessary and sufficient that all eigenvalues of matrix $\boldsymbol{A}$ are negative real or with a negative real part.

It might have been possible to start from system (3.31), perform Laplace transformation on this system, and solve. The results would have been identical.

3.3 Stability Analysis of Feedback Systems

Previously, it was shown that systems exhibit a different behaviour when they are in open loop or closed loop. The closed-loop response to a set point or a disturbance variation is influenced by the presence of actuators, measurement devices and, of course, controllers. To design a feedback control system, it is necessary to proceed to a stability study.

In open loop, the output $Y(s)$ is simply related to the control variable $U(s)$ and to the disturbance $D(s)$ by the relation

$$Y(s) = G_p(s)\,U(s) + G_d(s)\,D(s) \tag{3.35}$$

If the system is subjected to a known input, e.g. a step, the asymptotic behaviour will be found by using the final value theorem

$$\lim_{t\to\infty} y(t) = \lim_{s\to 0} s\,Y(s) \tag{3.36}$$

One already knows that when the process transfer function possesses positive real poles, or complex poles with a positive real part, the process is naturally unstable. The same reasoning can be made in closed loop (Fig. 2.19), the output $Y(s)$ being related to the set point $Y_r(s)$ and to the disturbance by the relation

$$Y(s) = G'_r(s)Y_r(s) + G'_d(s)D(s) \tag{3.37}$$

with the closed-loop transfer function relative to the set point

$$G'_r(s) = \frac{G_c(s)\,G_a(s)\,G_p(s)\,K_m}{1 + G_c(s)\,G_a(s)\,G_p(s)\,G_m(s)} \tag{3.38}$$

and the closed-loop transfer function relative to the disturbance

$$G'_d(s) = \frac{G_d(s)}{1 + G_c(s)\,G_a(s)\,G_p(s)\,G_m(s)} \tag{3.39}$$

Both closed-loop transfer functions $G'_r(s)$ and $G'_d(s)$ can be written as the ratio of two polynomials

$$\begin{aligned} G'_r(s) &= \frac{N'(s)}{D'(s)} \qquad \deg(N') \le \deg(D') \\ G'_d(s) &= \frac{N''(s)}{D''(s)} \qquad \deg(N'') \le \deg(D'') \end{aligned} \tag{3.40}$$

Conditions relying on the respective degrees of numerator and denominator express the fact that the system must be physically realizable (condition of physical realizability).

The closed-loop stability analysis will consist of the study of the positions in the complex plane of the poles of the closed-loop transfer functions, hence the zeros of denominators $D'(s)$ and $D''(s)$.

In general, the open-loop transfer function G_d relative to the disturbance has no common poles with G_c, G_a, G_p and G_m because these different transfer functions are obtained by identification or calculation in a real process.

If the transfer function G_d has common poles with G_c, G_a, G_p or G_m, a particular study must be realized to take it into account. If the transfer function G_d has unstable poles which are not cancelled by those of G_c, G_a, G_p or G_m, the closed-loop transfer function relative to the disturbance G'_d has the same poles and the feedback system is thus unstable for any disturbance.

Now study the case of the closed-loop transfer function $G'_r(s)$ relative to the set point. In this case, as the poles of $[G_c\, G_a\, G_p]$ are common with those of $[1 + G_c\, G_a\, G_p\, G_m]$ (which has only in addition the poles of G_m), it is equivalent to study the zeros of the characteristic equation denoted by

$$\begin{array}{l} 1 + G_c(s)\, G_a(s)\, G_p(s)\, G_m(s) = 0 \qquad \Longleftrightarrow \\ 1 + G_{ol}(s) = 0 \end{array} \tag{3.41}$$

where $G_{ol}(s)$ is the transfer function of the open loop, i.e. the product of all the transfer functions found in the loop of Fig. 2.19 (open must be considered as open between the measurement device and the comparator). When the characteristic equation has positive real zeros, or complex zeros with a positive real part (values in the right half s-plane) the closed-loop system is unstable.

Let us study the case of the closed-loop transfer function $G'_d(s)$ relative to the disturbance. We notice that its poles are those of the open-loop transfer function G_d, to which must be added the zeros of the characteristic equation. Thus it will be sufficient to complete the analysis of the characteristic equation with that of the transfer function G_d. In the case where the transfer function relative to the disturbance G_d has a positive pole or a pole with a positive real part, and that this pole is not common to G_c, G_a, G_m or G_p, this pole cannot be cancelled in closed loop and the system is unstable with respect to the disturbance, whatever the controller choice.

3.3.1 Routh-Hurwitz Criterion

The search of the zeros of the characteristic equation is far from evident without calculator or computer. Routh-Hurwitz criterion allows us to know analytically if at least one zero is situated in the right half s-plane which corresponds to an instability.

The characteristic equation is expressed in the form of a rational fraction where only the numerator will be considered

$$1 + G_c(s)\, G_a(s)\, G_p(s)\, G_m(s) = \frac{N(s)}{D(s)} \tag{3.42}$$

with:

$$N(s) = a_o s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n \tag{3.43}$$

where $a_0 > 0$ (else, the whole polynomial is multiplied by -1).
Property:

The Routh-Hurwitz criterion is a necessary and sufficient condition for the stability analysis: the number of roots of $N(s)$ that are positive real or complex with a positive real part is equal to the number of changes in sign of the first column of the associated array.

A polynomial with real coefficients is called Hurwitz if all its roots have a negative real part.

First search: if at least one coefficient a_i is negative, there exists at least a positive real zero or complex with a positive real part, and the polynomial $N(s)$ is called unstable.

Second search: if all a_i are positive, the following array is formed by $n+1$ rows:

row "s^n"	a_0	a_2	$a_4 \ldots$
row "s^{n-1}"	a_1	a_3	$a_5 \ldots$
row "s^{n-2}"	b_2	b_4	$b_6 \ldots$
row "s^{n-3}"	c_3	c_5	$c_7 \ldots$
$\ldots$			
row "s^1"	j_{n-1}		
row "s^0"	k_n		

where the first two rows are formed by the coefficients of the studied polynomial. The following rows of the array are related by the following relations:

first the row of "s^{n-2}":

$$b_2 = \frac{a_1 a_2 - a_0 a_3}{a_1} \quad b_4 = \frac{a_1 a_4 - a_0 a_5}{a_1} \quad b_6 = \frac{a_1 a_6 - a_0 a_7}{a_1} \quad \ldots$$

then the row of "s^{n-3}":

$$c_3 = \frac{b_2 a_3 - a_1 b_4}{b_2} \quad c_5 = \frac{b_2 a_5 - a_1 b_6}{b_2} \quad c_7 = \frac{b_2 a_7 - a_1 b_8}{b_2} \quad \ldots$$

then the row of "s^{n-4}":

$$d_4 = \frac{c_3 b_4 - b_2 c_5}{c_3} \quad d_6 = \frac{c_3 b_6 - b_2 c_7}{c_3} \quad \ldots$$

The array so formed is triangular, except for the last two rows with only one element per row.

Consider the elements of the first column of the array: a_0, a_1, b_2, c_3, …, k_n. If one of these elements is negative, at least one zero is positive real, or complex with a positive real part: the polynomial $N(s)$ is unstable.

Example 3.3: **Stability Analysis Using Routh-Hurwitz Criterion**

According to Fig. 3.8, let us consider a process with the following transfer function

$$G_p(s) = \frac{1}{(s+4)(s+6)} \tag{3.44}$$

measurement $G_m = 8$ and actuator

$$G_a(s) = \frac{1}{(s+1)} \tag{3.45}$$

Let us suppose that a simple proportional controller is used ($G_c = K_c$) as is

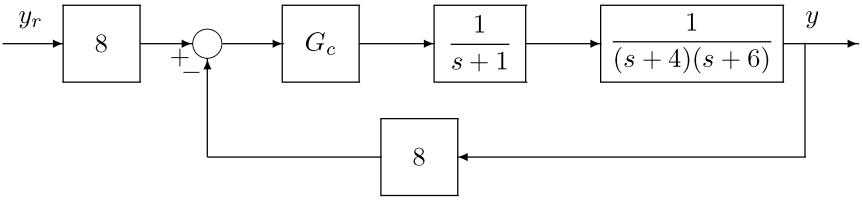


Figure 3.8: Block diagram for Example 3.3

often done for PID tuning (Sects. 4.5.1, 5.5). The characteristic equation is

$$1 + G_c G_a G_p G_m = 0 \quad \Longleftrightarrow \quad 1 + \frac{8\,K_c}{(s+1)(s+4)(s+6)} = 0$$
$$\text{or: } (s+1)(s+4)(s+6) + 8K_c = 0 \quad \Longleftrightarrow \quad s^3 + 11\,s^2 + 34s + 24 + 8K_c = 0 \tag{3.46}$$

All coefficients of this polynomial are positive, thus the second search phase must be performed.

Set the Routh-Hurwitz table:

row "s^3"	1	34
row "s^2"	11	$24 + 8K_c$
row "s^1"	$\frac{11 \times 34 - 1 \times (24 + 8K_c)}{11}$	0
row "s^0"	$24 + 8K_c$	0

Assuming K_c to be positive, the only element of the first column that is a potential problem is the coefficient of line "s^1" which must be positive. One obtains

$$11 \times 34 - 1 \times (24 + 8K_c) > 0 \quad \Longrightarrow \quad K_c < 43.75 \tag{3.47}$$

When the gain of the proportional controller is larger than the limit value 43.75, some roots of the characteristic equation become positive or have a positive real part, and the closed-loop system becomes unstable.

3.3.2 Root Locus Analysis

In the examples of the previously studied controller, it was shown that the controller gain influences the closed-loop response. The root locus (or Evans locus) is the set of the points of the complex plane described by the roots of the characteristic Eq. (3.41) when the controller gain varies between 0 and infinity. The values of the gain for which the roots are situated in the right half s-plane correspond to an unstable closed-loop system.

Nowadays, packages allow to easily find the roots of a polynomial, thus the zeros of the characteristic equation and thus allow us to easily obtain exact root loci. However, some indications are given in order to approximately draw these root loci:

- As the characteristic polynomial (3.43) has real coefficients, its roots are either real or a conjugate complex, thus the root locus is symmetrical with respect to the horizontal real axis,
- One begins by placing the poles of the open-loop transfer function G_{ol} (Eq. (3.41)) and they are denoted differently (by a circle in Fig. 3.9) from the future closed-loop poles (which will be denoted by a cross in Fig. 3.9),
- The asymptotic lines of the root locus are determined. These asymptotes are obtained by making s tend towards infinity. The number N of asymptotes is equal to the relative degree of the open-loop transfer function G_{ol}, i.e. the difference between the degrees of the denominator and the numerator of G_{ol}, or still the number of poles n_p of G_{ol} minus the number of zeros n_z.

  The trigonometric angles of the asymptotic lines are equal to

$$\text{Angles of the asymptotes} = \frac{\pm(2k+1)\pi}{N} \qquad k = 0, 1, \ldots, N-1 \tag{3.48}$$

  Note that an asymptote having an angle equal to π is the half real negative axis, i.e. a line starting from a given real value and tending towards $-\infty$.

  Let us denote the zeros of G_{ol} by z_i and its poles by p_i; G_{ol} can be rearranged as

$$G_{ol} = \frac{K_c}{s^{n_p - n_z} + (p_1 + \ldots + p_{n_p} - z_1 - \ldots - z_{n_z}) s^{n_p - n_z - 1} + \ldots} \tag{3.49}$$

  where K_c is the controller gain. As the characteristic equation corresponds to $G_{ol}(s) = -1$, and we are interested in the behaviour when s tends towards infinity, considering the two higher-degree terms gives an approximation of the equation giving a unique value of s as

$$\left(s + \frac{(p_1 + \ldots + p_{n_p}) - (z_1 + \ldots + z_{n_z})}{n_p - n_z} \right)^{n-m} = 0 \tag{3.50}$$

  The position of the point of the real axis which is the intersection of the asymptotes is thus deduced

$$\begin{aligned} s &= -\frac{(p_1 + \ldots + p_{n_p}) - (z_1 + \ldots + z_{n_z})}{n_p - n_z} \\ &= -\frac{\text{sum of poles} - \text{sum of zeros}}{\text{number of poles} - \text{number of zeros}} \end{aligned} \tag{3.51}$$

- The number of branches is equal to the degree of the characteristic equation, i.e. the number of poles for a proper transfer function.

  The branches of the root locus start (when $K_c \to 0$) at the poles of G_{ol} and end (when $K_c \to \infty$) at the zeros of G_{ol}. If the number of closed-loop poles is equal to the number n_p of open-loop poles, the number of branches ending at finite zeros is equal to the number n_z of open-loop zeros. The other $n_p - n_z$ branches tend towards the asymptotes at infinity (the zeros are called zeros at infinity). The point where the branches break away

corresponds to multiple roots of the characteristic equation and is a solution of the following equation

$$\frac{dK_c}{ds} = 0 \tag{3.52}$$

where K_c has been expressed as a function of s deduced from the characteristic Eq. (3.42).

- It is possible to determine other characteristics in order to obtain more precisely the branches of the root locus (Ogata, 1997).

Example 3.4: **Root Locus**

Consider again Example 3.3. According to Fig. 3.8, the process has the following transfer function

$$G_p(s) = \frac{1}{(s+4)(s+6)} \tag{3.53}$$

measurement: $G_m = 8$ and actuator

$$G_a(s) = \frac{1}{(s+1)} \tag{3.54}$$

The open-loop transfer function is thus equal to

$$G_{ol} = G_c G_a G_p G_m = G_c \frac{8}{(s+1)(s+4)(s+6)} \tag{3.55}$$

Its poles are negative so that the system is stable in open loop. We wish to study the influence of the controller gain and we assume that the controller is a simple proportional one: $G_c = K_c$. The characteristic equation is thus

$$1 + G_{ol} = 0 \quad \Longleftrightarrow \quad (s+1)(s+4)(s+6) + 8K_c = 0 \tag{3.56}$$

Its solutions depend on the values of the proportional controller gain.

According to the previous rules, we notice that G_{ol} presents three poles and no zero. It follows that the root locus (Fig. 3.9) will have three asymptotes at $\pi/3$, π and $-\pi/3$ which intersect at the real axis point of abscissa: $-(1+4+6)/3 \approx -3.66$.

The branches start at the poles of G_{ol}, i.e. at the points on the real axis of abscissae -1, -4 and -6.

The two branches having a complex part start at -1 and -4 and break away at the point that we find from Eq. (3.56) by setting

$$\frac{dK_c}{ds} = 0 \qquad \text{with:} \quad K_c = -\frac{(s+1)(s+4)(s+6)}{8} \tag{3.57}$$

which gives the two roots -5.12 and -2.21. Only the root -2.21 conveys which corresponds to the value $K_c = 1.026$, at which we begin to obtain complex closed-loop poles (or roots of the characteristic equation). The other root corresponds to a negative value of K_c.

The third branch on the real axis starts from -6 and is directed towards $-\infty$.

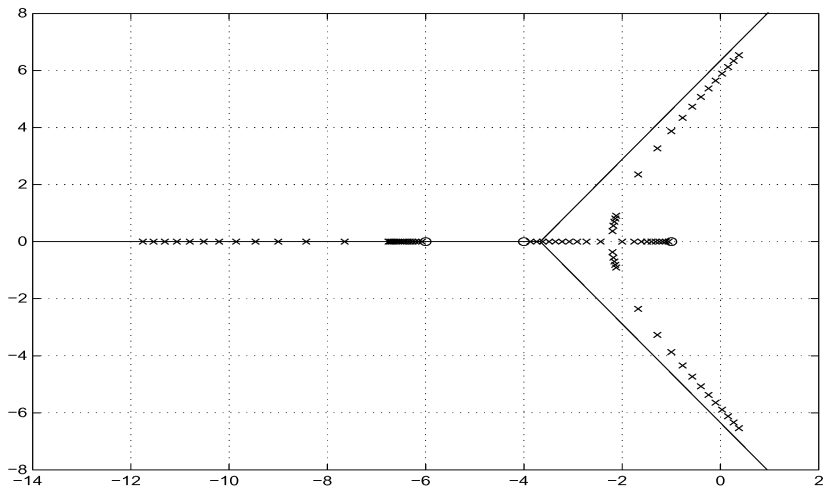


Figure 3.9: Root locus diagram of a third-order open-loop transfer function. The poles of the open-loop transfer function and the closed-loop poles are respectively represented by "o" and by "×". The asymptotes are represented by continuous lines

In Fig. 3.9, the root locus (symbol "×") was obtained for gain values included between 0.1 and 1.5 by increments of 0.1, then between 5 and 60 by increments of 5. Moreover, the figure displays the open-loop poles (here, points -1, -4 and -6 of real axis). For gain values smaller than 1.026, the three closed-loop roots are real and two of them become conjugate complex beyond this value. Moreover, when the gain becomes larger than around 40, these conjugate complex roots have their real part positive: in closed loop, the system becomes unstable for high values of the controller gain although it is stable in open loop.

Example 3.5: **Root Locus**

According to Fig. 3.10, let us assume that a process has the following transfer function

$$G_p(s) = \frac{4(2s+1)}{(5s+1)(25s^2+5s+1)} \tag{3.58}$$

and that $G_m = 1$ and $G_a = 1$.
The open-loop transfer function is thus equal to

$$G_{ol} = G_c G_a G_p G_m = G_c \frac{4(2s+1)}{(5s+1)(25s^2+5s+1)} \tag{3.59}$$

Besides G_c, one pole is negative real ($s = -0.2$) and the two other poles have a negative real part ($s = -0,1 \pm 0,1732i$), thus the system is open-loop stable. Moreover, G_{ol} has a negative real zero ($s = -0.5$). To obtain the root locus, we

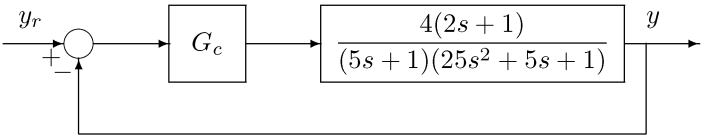


Figure 3.10: Block diagram for Example 3.5

assume that the controller is simply proportional: $G_c = K_c$. The characteristic equation takes the form

$$1 + G_{ol} = 0 \quad \Longleftrightarrow \quad (5s+1)(25s^2+5s+1) + 4K_c(2s+1) = 0 \qquad (3.60)$$

G_{ol} presents three poles and a negative zero. The relative degree of the open-loop transfer function G_{ol} is equal to the difference between the degree of the denominator and that of the numerator, i.e. 2. The root locus (Fig. 3.11) will have two asymptotes at $\pi/2$ and $-\pi/2$, which intersect at the point on the real axis of abscissa: $-((-0.2 - 2 \times 0.1732) - (-0.5))/(3-1) \approx 0.05$.

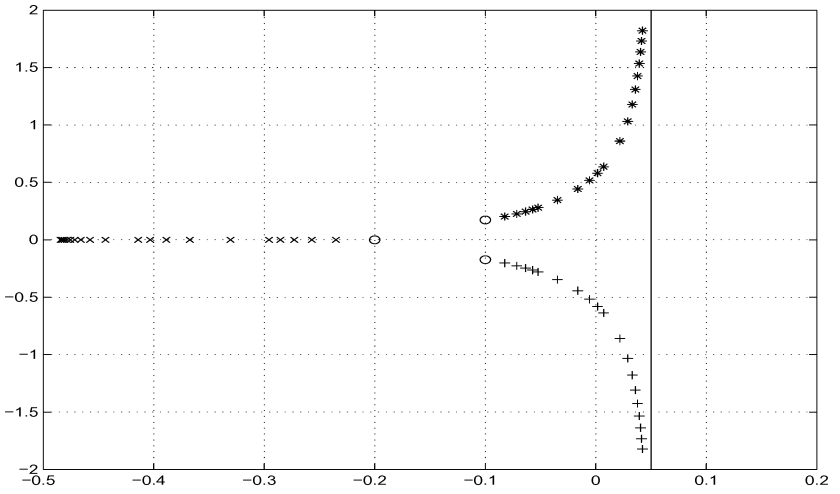


Figure 3.11: Root locus for open-loop transfer function (3.59) of relative degree 2. The poles of the open-loop transfer function and of the closed-loop transfer function are respectively represented by "o" and by "×". The asymptotes are represented by continuous lines

The root locus has three branches starting at the poles of G_{ol}: one of the branches corresponding to the real roots starts from the pole of G_{ol} on the real axis at abscissa -0.2 and ends at the zero of G_{ol} at abscissa -0.5; the two other branches correspond to the conjugate complex roots and start at the points of the complex plane of affixes $-0.1 \pm 0.1732i$.

Figure 3.11 gives the root locus (symbol "×") obtained for gain values successively comprised between 0.1 and 0.5 in increments of 0.1, then between 1 and 5 in increments of 1, then between 10 and 50 in increments of 5. Moreover, the figure shows the open-loop poles (here, point −0.2 of the horizontal real axis and points $-0.1 \pm 0.1732i$ in the complex plane). It appears clearly in the figure that the root loci originate from the open-loop poles. When the gain is larger than about 3.77, two roots have their real part positive and the closed-loop system becomes unstable.

3.3.3 Frequency Method

The frequency method consists of searching the frequency ω and the maximum gain K_{cm} of the proportional controller such that the characteristic equation is equal to zero for these values (see Chap. 5). The analysis is realized in the complex plane and it is sufficient to set

$$s = j\,\omega \tag{3.61}$$

Example 3.6: **Stability Analysis in Frequency Domain**

Take again the previous Example 3.3 where the open-loop transfer function is equal to

$$G_{ol} = G_c G_a G_p G_m = G_c \frac{8}{(s+1)(s+4)(s+6)} \tag{3.62}$$

with $G_c = K_c$ (the controller is set in proportional mode) and the characteristic equation to be solved becomes

$$\begin{array}{l} (s+1)\,(s+4)\,(s+6) + 8\,K_c = 0 \Longleftrightarrow \\ (j\,\omega+1)\,(j\,\omega+4)\,(j\,\omega+6) + 8\,K_c = 0 \Longleftrightarrow \\ (-11\,\omega^2 + 24 + 8\,K_c) + j\,(-\omega^3 + 34\,\omega) = 0 \Longleftrightarrow \\ \left\{ \begin{array}{ll} -11\omega^2 + 24 + 8\,K_{cm} & = 0 \\ -\omega^3 + 34\,\omega & = 0 \end{array} \right. \end{array} \tag{3.63}$$

from which we draw the frequency $\omega = 5.831$ radians/time unit and the maximum gain of the proportional controller $K_{cm} = 43.75$, beyond which the system becomes unstable. For this limit gain, the system subjected to a set point step-like variation would present an output showing sustained oscillation.

References

R. Aris and N.R. Amundson. An analysis of chemical reactor stability and control I, II, III. *Chem. Eng. Sci.*, 7:121–155, 1958.

J.S. Berger and D.D. Perlmutter. Chemical reactor stability by Liapunov's direct method. The effect of feedback control on chemical reactor stability. *AIChE. J.*, 10(2):233–245, 1964.

B. d'Andrea Novel and M. Cohen de Lara. *Commande Linéaire des Systèmes Dynamiques.* Masson, Paris, 1994.

R. Gilmore. *Catastrophe Theory for Scientists and Engineers.* Wiley, New York, 1981.

J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields.* Springer-Verlag, New York, 1986.

P.M. Hidalgo and C.B. Brosilow. Nonlinear model predictive control of styrene polymerization at unstable operating points. *Comp. Chem. Eng.*, 14(4/5): 481–494, 1990.

D.M. Himmelblau and K.B. Bischoff. *Process Analysis and Simulation.* Wiley, New York, 1968.

K. Ogata. *Modern Control Engineering.* Prentice Hall, Englewood Cliffs, New Jersey, 1997.

T.S. Parker and L.O. Chua. *Practical Numerical Algorithms for Chaotic Systems.* Springer-Verlag, New York, 1989.

L. Pellegrini, G. Biardi, and M.G. Grottoli. Determination of the region of asymptotic stability for a CSTR. *Comp. Chem. Eng.*, 12(2/3):237–241, 1988.

C. Storey. Liapunov stability. In N. Munro, editor, *Modern Approaches to Control System Design*, volume 9, chapter 16, pages 325–337. Institution of Electrical Engineers, London, 1979.

A. Uppal, W.H. Ray, and A.B. Poore. On the dynamic behaviour of continuous stirred tank reactors. *Chem. Eng. Sci.*, 29:967–985, 1974.

C. Van Heerden. Autothermic process. *Ind. Eng. Chem.*, 45:1242–47, 1953.

M. Vidyasagar. *Control Systems Synthesis: A Factorization Approach.* MIT Press, Cambridge, Massachusetts, 1985.

R.B. Warden, R. Aris, and N.R. Amundson. An analysis of chemical reactor stability and control – IX : Further investigations into the direct method of Lyapunov. *Chem. Eng. Sci.*, 19(149):173–190, 1964.

Chapter 4

Design of Feedback Controllers

Until this chapter, the problem of getting the best values for the controller parameters so as to get the "best" possible response was not our objective. In this chapter, solutions will first concern classical PID controllers. The important problems are:

- The choice of the controller type.
- The tuning of the controller parameters.
- The performance criteria to be used.

After the PID controller, more sophisticated control methods will be explained such as internal model control, pole-placement control and linear quadratic control.

4.1 Performance Criteria

The controller role for the closed-loop system is to guarantee that the response presents suitable dynamic and steady-state characteristics (Kestenbaum et al., 1976).

The following criteria can be cited:

- The controller must be able to maintain the controlled variable to its set point.
- The closed-loop system must be asymptotically stable and present a satisfactory performance in a large range of frequencies.
- The influence of disturbances must be minimized.
- The responses to set point variations must be fast and smooth.
- An excessive control must be avoided (the control variable u must not undergo too frequent large and fast variations).
- The control system must be robust: it must be insensitive to process variations and modelling errors.

In reality, all these objectives cannot be simultaneously realized and the control system (Fig. 4.1) results from a compromise. For industrial applications, the robustness is particularly important.

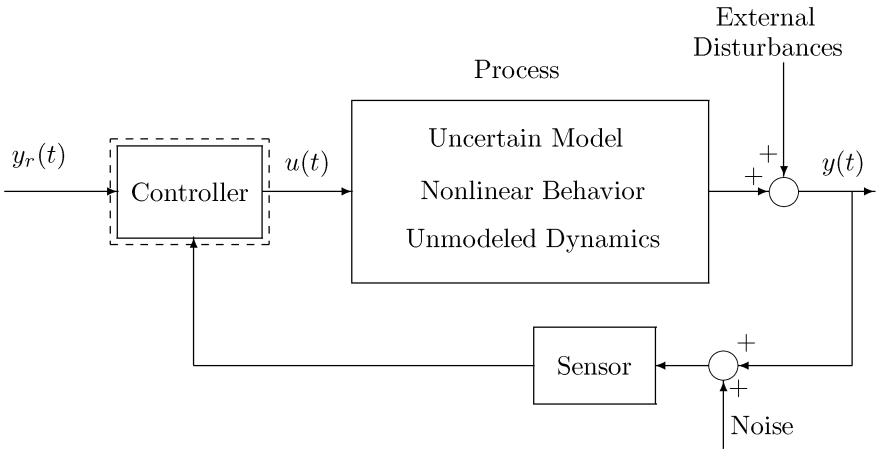


Figure 4.1: Summary of the problems of a controller design for process control

The more frequently used criteria (cf. Fig. 1.29) are:

- The overshoot.
- The rising time.
- The stabilization time.
- The decay ratio.
- The oscillation frequency of the transient.

Different methods are possible:

- "Direct synthesis" method.
- Internal model control.
- Pole-placement.
- Optimal control.
- Tuning relations.
- Frequency response techniques.
- Computer simulation based on knowledge models.
- On-site tuning.

In the context of this chapter, the first five methods are based on continuous transfer function models or Laplace polynomials; frequency response techniques can be used for any linear model. Optimal control based on state-space models will be studied in Chap. 14. Computer simulation allows us to use any type of model, in particular knowledge models based on fundamental principles, but is longer in this case and needs a lot of human investment.

4.2 Transient Response Characteristics

The characteristics of the transient response depends on the position of the closed-loop transfer function poles. The dominant poles which are the closest to origin are retained. When only a given pair of complex poles is retained, the reduced system has the characteristics of a second-order system with the following closed-loop transfer function

$$G(s) = \frac{\omega_n^2}{\omega_n^2 + 2\,\zeta\,\omega_n s + s^2} \tag{4.1}$$

with damping factor ζ and natural frequency ω_n. On the left half s-plane, corresponding to stable poles, different features can be drawn (Fig. 4.2) for a step response:

- The half straight lines of constant overshoot corresponding to relation (1.157) hence
$$y = \pm\frac{\sqrt{1-\zeta^2}}{\zeta}x \qquad \text{with: } \zeta = \text{constant} < 1 \tag{4.2}$$
the damping factor ζ must not be too low; the first bisecting line corresponds to value $\zeta = \sqrt{2}/2$.
- The half circle of constant rising time corresponding to relation (1.160) hence
$$t_m \approx \frac{2.5}{\omega_n} \approx \text{constant} \Longrightarrow x^2 + y^2 \approx \omega_n^2 \qquad \text{with: } x < 0 \tag{4.3}$$
- The straight line of constant stabilization time t_s corresponding to relation (1.162) hence
$$t_s \approx \frac{4.6}{\zeta\omega_n} \approx \text{constant} \Longrightarrow x = -\zeta\omega_n \approx -\frac{4.6}{t_s} \tag{4.4}$$

Figure 4.2 allows us to evaluate the influence of the pole position on these three frequently used criteria: the hatched zone must be avoided.

Moreover, in general, the real part of the poles must not be too small (its absolute value must not be too large), because in the opposite case, it would correspond to a very fast response but would probably have the consequence of saturating the manipulated variable.

4.3 Performance Criteria for Design

Denoting as an error the difference $e(t) = y(t) - y_r(t) = (\text{output} - \text{set point})$, the criteria dealing with the integral of an error function take into account the global nature of the process response. Several criteria exist:

- Integral of the square of the error (ISE)
$$\text{ISE} = \int_0^\infty e^2(t)\,dt \tag{4.5}$$

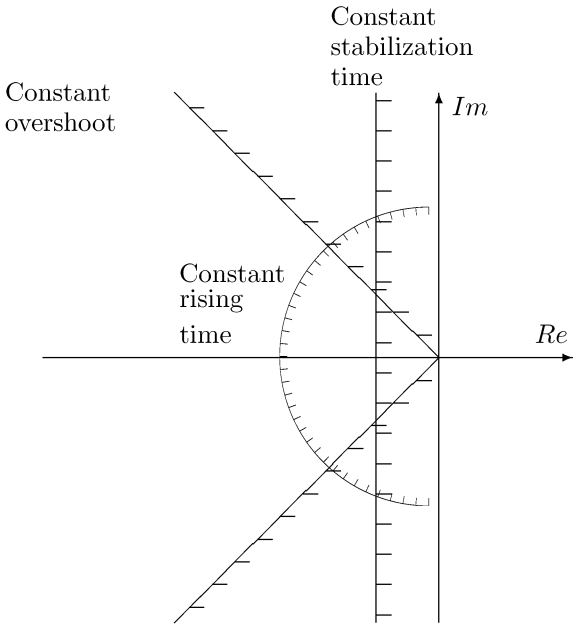


Figure 4.2: Loci of constant overshoot, rising time and stabilization in the s-plane. The hatched zone is not convenient for the poles

- Integral of the absolute value of the error (IAE)

$$\text{IAE} = \int_0^\infty |e(t)|\, dt \tag{4.6}$$

- Integral of the absolute value of the error weighted by time (ITAE)

$$\text{ITAE} = \int_0^\infty t\, |e(t)|\, dt \tag{4.7}$$

The problem is then to choose the controller type and its parameters so as to minimize one of the previous criteria. The criteria can be ordered with respect to their own characteristics:

- Use of IAE gives well-damped systems; for a second-order system, the resulting damping factor will be around 0,7 (Shinners, 1992) nearly as for ITAE. With ISE, it would be around 0,5. ISE is not very sensitive to parameter variations, as opposed to ITAE which is sensitive.
- To suppress large errors (numbers > 1), ISE is better than IAE because the error term intervenes by its square.
- To suppress small errors (numbers < 1), IAE is better than ISE.
- To suppress errors which persist a long time, ITAE is the best criterion as the t term amplifies small persisting errors. This is often the preferred criterion because it offers more security. In general, it will produce smaller overshoots and oscillations.

Graham and Lathrop (1953) have searched the best closed-loop transfer functions with respect to the ITAE criterion. Indeed, it is possible to consider several types of transfer functions G_{cl}:

- When the numerator is a constant, the transfer function has no zeros: this corresponds to an error of position which is zero for a step input. The denominator will take one of the forms given by the top array of Table 4.1.
- When the numerator degree is 1, the transfer function possesses one zero. It will be designed to ensure a zero position error and a zero velocity error for a ramp trajectory. The denominator will take one of the forms of the middle array of Table 4.1.
- When the numerator degree is 2, the transfer function possesses two zeros. It will ensure a zero acceleration error for parabolic trajectories. The denominator will take one of the forms of the bottom array of Table 4.1.

It must be noted that the higher the demand with respect to the transfer function G_{cl}, the more zeros it possesses and the more its transient behaviour can itself deteriorate (Fig. 4.3).

Table 4.1: Denominators of the optimal transfer functions for the ITAE criterion and for systems with zero position error, zero velocity error or zero acceleration error

Denominator of transfer function G_{cl} optimal for ITAE
For a system with zero position error $G_{cl}(s) = \omega_0^n/(s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + \omega_0^n)$
$s + \omega_0$
$s^2 + 1.4\omega_0 s + \omega_0^2$
$s^3 + 1.75\omega_0 s^2 + 2.15\omega_0^2 s + \omega_0^3$
$s^4 + 2.1\omega_0 s^3 + 3.4\omega_0^2 s^2 + 2.7\omega_0^3 s + \omega_0^4$
$s^5 + 2.8\omega_0 s^4 + 5.0\omega_0^2 s^3 + 5.5\omega_0^3 s^2 + 3.4\omega_0^4 s + \omega_0^5$
$s^6 + 3.25\omega_0 s^5 + 6.6\omega_0^2 s^4 + 8.6\omega_0^3 s^3 + 7.45\omega_0^4 s^2 + 3.95\omega_0^5 s + \omega_0^6$
For a system with zero velocity error $G_{cl}(s) = (a_1 s + \omega_0^n)/(s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + \omega_0^n)$
$s^2 + 3.2\omega_0 s + \omega_0^2$
$s^3 + 1.75\omega_0 s^2 + 3.25\omega_0^2 s + \omega_0^3$
$s^4 + 2.41\omega_0 s^3 + 4.93\omega_0^2 s^2 + 5.14\omega_0^3 s + \omega_0^4$
$s^5 + 2.19\omega_0 s^4 + 6.50\omega_0^2 s^3 + 6.30\omega_0^3 s^2 + 5.24\omega_0^4 s + \omega_0^5$
$s^6 + 6.12\omega_0 s^5 + 13.42\omega_0^2 s^4 + 17.16\omega_0^3 s^3 + 14.14\omega_0^4 s^2 + 6.76\omega_0^5 s + \omega_0^6$
For a system with zero acceleration error $G_{cl}(s) = (a_2 s^2 + a_1 s + \omega_0^n)/(s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + \omega_0^n)$
$s^3 + 2.97\omega_0 s^2 + 4.94\omega_0^2 s + \omega_0^3$
$s^4 + 3.71\omega_0 s^3 + 7.88\omega_0^2 s^2 + 5.93\omega_0^3 s + \omega_0^4$
$s^5 + 3.81\omega_0 s^4 + 9.94\omega_0^2 s^3 + 13.44\omega_0^3 s^2 + 7.36\omega_0^4 s + \omega_0^5$
$s^6 + 3.93\omega_0 s^5 + 11.68\omega_0^2 s^4 + 18.56\omega_0^3 s^3 + 19.30\omega_0^4 s^2 + 8.06\omega_0^5 s + \omega_0^6$

In Table 4.1, the value of ω_0 is not given. It depends on the desired dynamics for the system; in some cases, it can be determined with respect to constraints existing on the actuator: $u_{\min} \leq u(t) \leq u_{\max}$.

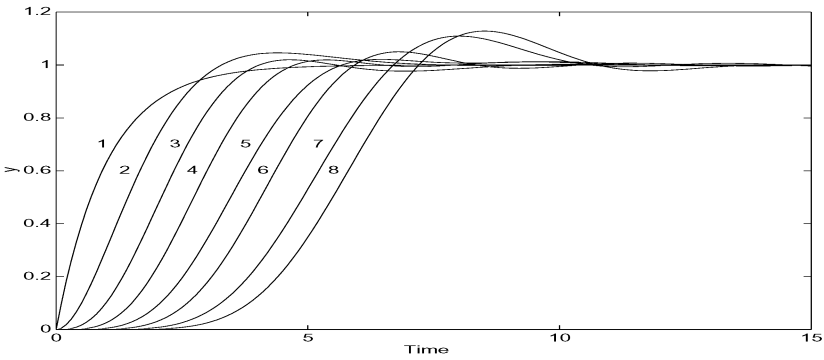


Figure 4.3: Responses to a step input for optimal transfer functions with respect to the ITAE criterion in the case of a zero position error. The number on the figure indicates the transfer function order

4.4 Choice of PID Controller

4.4.1 General Remarks

Up to now, a given number of generalities can be drawn:

- Proportional action:
  - Accelerates the process response by gain increase.
  - Produces a steady-state deviation except for processes which have an integrator term $(1/s)$ in their transfer function. This offset decreases when the proportional gain increases.
- Integral action:
  - Eliminates the steady-state deviation.
  - This elimination is done in general at the expense of larger deviations.
  - Responses are sluggish, with long oscillations
  - The increase of gain K makes the behaviour more oscillatory and can lead to instabilities.
- Derivative action:
  - Anticipates future errors.
  - Introduces a stabilizing effect in the closed-loop response.

4.4.2 Recommendations

Simple Rules

1) If possible, use a P controller: if the steady-state deviation is tolerable, or if the process contains an integral term. Example: gas pressure control or level control.
2) If the P controller is unacceptable, use a PI: if the steady-state deviation is too large. Example: flow rate control. In this case, the response is fast and the slowing induced by integral action is unimportant. Aström and Hägglund (1988) recommend using a PI controller for processes that have first-order dynamics, e.g. level control in a tank.
3) In other cases, use a PID: the closed-loop response will be faster and the controller will be more robust. Example: temperature control, composition control, processes with capacities in series. Aström and Hägglund (1988) recommend using PID controllers for processes having second-order dynamics, which sometimes may be difficult to detect, or having time constants of different orders of magnitude. Owing to the derivative action, the gain can be limited.
4) Typical systems posing serious problems for PID are:

- Systems with time delay.
- Systems with oscillatory modes.
- Systems with large parameter variations.
- Systems for which a quality variable should be controlled.

More Detailed Discussion

This advice is valid for use in the case where there is no available process model; they do not constitute instructions for use without reservation and will always need to be used with care.

Flow rate control. Control feedbacks for flow rate and liquid pressure are characterized by fast responses so that the delays are, in general, negligible. The sensor and the pneumatic transmission lines can introduce a time delay. Disturbances are, in general, high-frequency noise, which makes the derivative action unusable. PI controllers are frequently used.

Liquid level control. Integral action is not necessary if a small deviation (around 5%) is tolerated at the level. If integral action is used, high gains can be chosen due to the integrating nature of the process. In general, the derivative action is not used. In many cases, a buffer tank is used to avoid level fluctuations in the plant. In this case, the outlet flow rate of the tank must be as stable as possible and the controller will have to be carefully tuned.

When heat transfer occurs in the tank (reboiler, evaporator), its operating model is more complicated and the controller will be different.

Gas pressure control. If the gas is in equilibrium with a liquid, gas pressure control is difficult. Here, pressure control is considered for a gas alone. The tank (or the pipe) partially controls itself: if the pressure inside the system becomes too high, the feed flow rate decreases, and vice-versa. In general, PI controllers are used with a small integral action (τ_I large). Often, tank volumes are small so that the residence times are low with respect to the rest of the process and derivative action is not necessary.

Temperature control. Temperature control problems are complicated and of a large variety with respect to the considered system. Frequently, because of occurring time delays, the gain must not be too large to avoid instability. PID controllers are often used to have a faster response than with a PI and to stabilize the process.

Composition control. Several points are common with temperature control, but two additional factors intervene:

– The measurement noise is more important.
– Time delays can be very large (example: chromatographs).

4.5 PID Controller Tuning

In many cases, the technician or engineer faces an existing plant and must do on-site tuning.

4.5.1 Tuning by Trial and Error

A typical on-site tuning of a PID controller (later, the connection with the Bode stability criterion (Sect. 5.5) and Ziegler-Nichols tuning will appear) is done as follows:

- Stage 1: The controller is set in proportional mode only (it suffices to take τ_I maximum and τ_D minimum).
- Stage 2: A low value of the gain is chosen and the controller is set in automatic mode. A step set point variation is operated.
- Stage 3: The gain is increased by small increments until a sustained oscillation of period T_u is obtained (corresponding to the ultimate gain K_{cu}).
- Stage 4: The gain is reduced by a factor of 2.
- Stage 5: τ_I is decreased by small increments until a sustained oscillation is again obtained. τ_I is set to three times this value.
- Stage 6: τ_D is increased until sustained oscillation is obtained. τ_D is set to one third of this value.

During this operation, it is necessary to avoid saturating the controller output. Otherwise, oscillations can occur at gain values lower than the ultimate gain K_{cu}. In principle, when the gain is lower than the ultimate gain, the closed-loop response $u_a(t)$ (the controller output) is underdamped or weakly oscillating. When the gain is larger than the ultimate gain, the system is unstable and the

response is theoretically unbounded. As a matter of fact, saturation will occur very often.

Conclusion: the ultimate gain K_{cu} is the highest value of the controller gain for which the system is closed-loop stable, the controller being only in proportional mode.

Drawbacks: this procedure can be unsafe, e.g. it can lead to reactor runaway. Moreover, it cannot be applied to processes which are unstable in open loop. On the other hand, some processes have no ultimate gain: e.g. a process perfectly modelled by a first- or second-order transfer function and having no time delay. Nevertheless, the rules which are given clearly show the influence of each action and provide a line of conduct: when the tendency to sustained oscillations is detected, a safety margin is adopted for the controller parameters.

4.5.2 Sustained Oscillation Method

The method presented in the previous section is only a variant of the continuous cycling method by Ziegler and Nichols (Ziegler and Nichols, 1942). This method consists of determining as previously the ultimate gain. The sustained oscillation period is called the ultimate period T_u. Then the tuning recommended by Ziegler and Nichols can be used (Table 4.2). Furthermore, this only constitutes a first tuning which must be refined.

Table 4.2: PID tuning recommended by Ziegler and Nichols

Controller	K_c	τ_I	τ_D
P	$0.5\,K_{cu}$		
PI	$0.45\,K_{cu}$	$T_u/1.2$	
PID	$0.6\,K_{cu}$	$T_u/2$	$T_u/8$

It can be noticed that the gain of the PI controller is lower than that of the P controller, and that of the PID controller is the highest, thus corresponding to the general tendencies of the integral and the derivative actions.

Indeed, these empirical settings based on a decay ratio equal to 0.25 are not necessarily the best; they cause a non-negligible overshoot, as a decay ratio equal to 0.25 gives a damping factor $\zeta = 0.22$, which is too low. Different settings are proposed (Perry, 1973) for PID controllers in Table 4.3 which produce less overshoot but may, however, not produce the foreseen results.

The tuning "without overshoot" corresponds of course to the lowest controller gain, but a small overshoot is frequently tolerated.

It may happen that one type of tuning is convenient for a set point variation but is less convenient when the system is subjected to a disturbance.

Table 4.3: Tuning of a PID giving less overshoot than Ziegler and Nichols tuning. Note that these recommendations by Perry (1973) do not always provide the expected results

Controller with	K_c	τ_I	τ_D
Light overshoot	$0.33\,K_{cu}$	$T_u/2$	$T_u/3$
No overshoot	$0.2\,K_{cu}$	$T_u/2$	$T_u/3$

4.5.3 Relay Oscillation Method

The method of sustained oscillation in the neighbourhood of instability is favourably replaced by the observation that a relay interposed in the loop before the actuator (Fig. 4.4) allows to generation of a control signal having the shape of a square wave, which leads to an oscillating measured output with the critical or ultimate period T_c. Indeed, the relay works as an on-off controller. The only influence of the proportional controller of Fig. 4.4 is on the relay excitation signal. Consequently, its gain is chosen to be equal to 1 (as if the proportional controller was absent) or a value which does not prevent the relay from oscillating. Moreover, the output oscillation amplitude is proportional to the relay amplitude and can be chosen to be small so as to little perturb the process. Thus this method is clearly more admissible in a real process than the trial and error tuning method, as the system continues to run in closed loop in a small neighbourhood of its set point (Aström and Hägglund, 1988; Voda, 1996). It is possible to use a relay without or with hysteresis, the latter being less sensitive to the noise and therefore likely to make a relay without hysteresis swing.

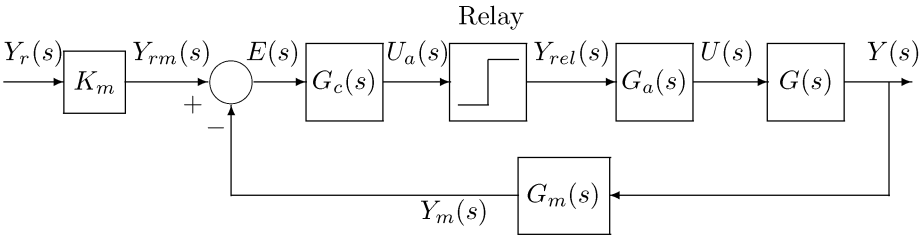


Figure 4.4: Use of a relay to provoke system oscillations for tuning (Aström and Hägglund, 1988). Note that in the presence of the relay, the proportional controller G_c has a gain of 1 so that the controller has no influence

Consider the case of a relay with hysteresis (also called dead zone relay). In the presence of the relay, the process output enters in a limit cycle that is to be characterized with respect to its amplitude and frequency (Coughanowr and Koppel, 1985). The error signal $e(t) = y_{rm}(t) - y_m(t)$ (Fig. 4.5) is approximated

by a sinusoidal function equal to

$$e(t) \approx A \sin(\omega t) \tag{4.8}$$

A perfectly sinusoidal error signal would correspond exactly to an ellipsoidal

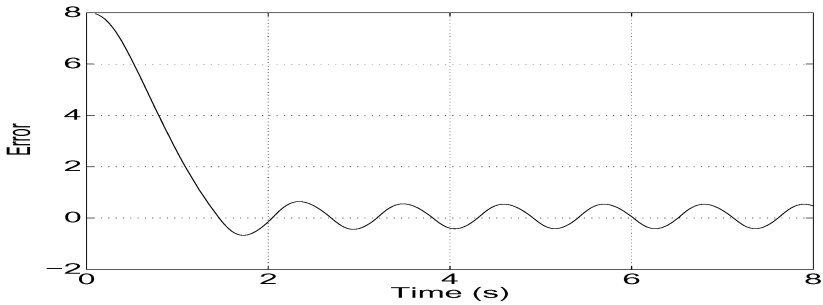


Figure 4.5: Error signal.

limit cycle.

Consider a relay with hysteresis whose state changes either by passing from a value $y_{rel,max}$ to a value $y_{rel,min}$ when the error signal which excites it crosses a small positive value ϵ by becoming lower than ϵ or by passing from a value $y_{rel,min}$ to a value $y_{rel,max}$ when the error signal crosses the value $(-\epsilon)$ by becoming larger than $(-\epsilon)$. The action of this relay can be symbolized by Fig. 4.6.

With respect to the error signal $e(t)$, the rectangular wave of the relay symbolized by Fig. 4.6 is a signal in advance by $1/\omega \arcsin(\epsilon/A)$ corresponding to the hysteresis. This rectangular wave $y_{rel}(t)$ of period $T = 2\pi/\omega$, can be described by means of the periodic function $f(t)$ with no advance

$$y_{rel}(t) = f(t + 1/\omega \arcsin(\epsilon/A)) \text{ , with: } f(t) = \begin{cases} y_{rel,max} & \text{if: } 0 < t < \dfrac{T}{2} \\ y_{rel,min} & \text{if: } \dfrac{T}{2} < t < T \\ f(t+T) = f(t) & \forall\, t \end{cases} \tag{4.9}$$

This function $f(t)$ can be expanded as a Fourier series as

$$f(t) = (y_{max} - y_{min}) \sum_{k=0}^{\infty} \frac{2}{(2k+1)\pi} \sin((2k+1)\,\omega\, t) \tag{4.10}$$

and the rectangular wave of the relay becomes

$$y_{rel}(t) = (y_{max} - y_{min}) \sum_{k=0}^{\infty} \frac{2}{(2k+1)\pi} \sin((2k+1)\,\omega\,(t + 1/\omega \arcsin(\epsilon/A))) \tag{4.11}$$

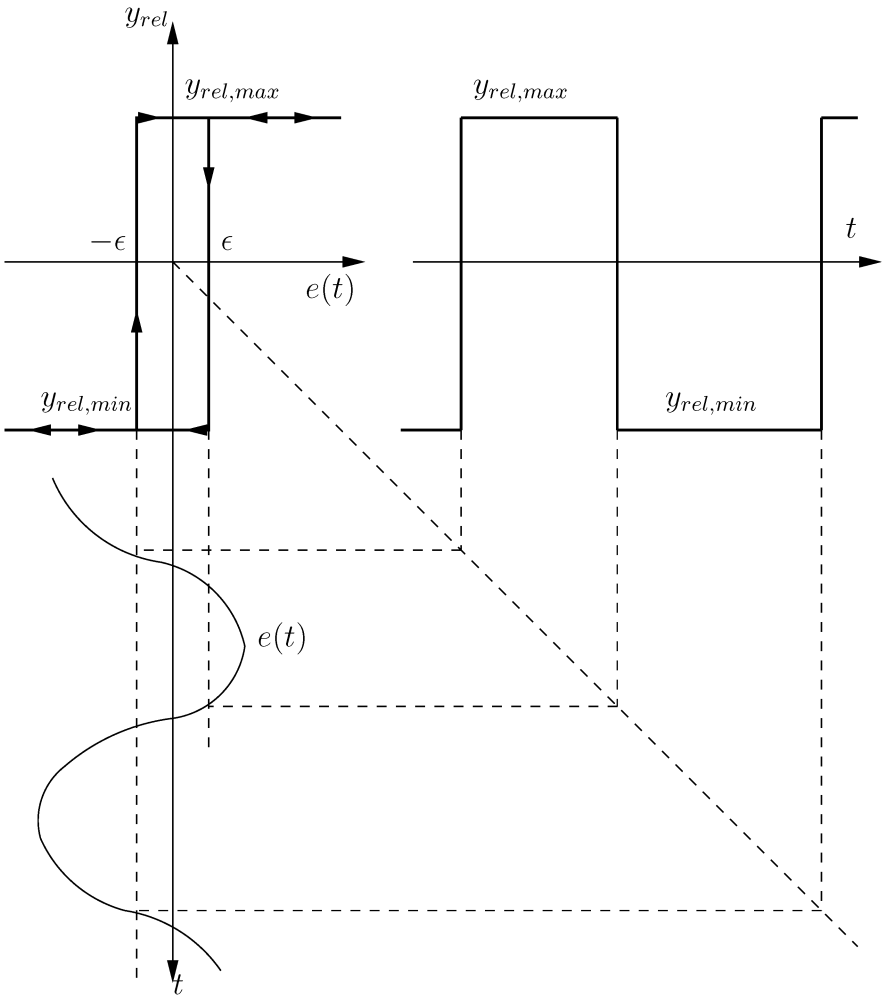


Figure 4.6: Action of relay with hysteresis excited by the error signal $e(t)$

which can be approximated by its first harmonics, as the influence of the other harmonics is weighted by lower coefficients and is largely filtered by the other elements of the feedback loop

$$y_{rel}(t) \approx (y_{max} - y_{min}) \frac{2}{\pi} \sin(\omega\,(t + 1/\omega \arcsin(\epsilon/A))) \tag{4.12}$$

Set the relay amplitude $a_u = (y_{max} - y_{min})/2$. Thus, the relay acts essentially through its first harmonics of amplitude $4a_u/\pi$ on the closed-loop process whose measured output oscillates with an amplitude a_y and a period T_c according to a limit cycle (Fig. 4.5). The relay which has the error $e(t)$ as an input and $y_{rel}(t)$ as an output is described by a nonlinear function, which is indeed a pure

gain $G_{rel}(a_u)$ depending on the amplitude a_u of the rectangular wave of the relay

$$G_{rel} = \frac{4a_u}{\pi a_y} \tag{4.13}$$

The condition of oscillation of the limit cycle is

$$G_{rel}(a_u)\, G_{proc}(j\omega_c) = -1 \tag{4.14}$$

with: $G_{proc}(s) = G_a(s)G(s)G_m(s)$, the critical pulsation ω_c being equal to

$$\omega_c = \frac{2\pi}{T_c} \tag{4.15}$$

In chapter 5, ω_c is called the phase crossover frequency ω_ϕ and T_c is denoted by the ultimate period T_u.

From the limit cycle condition, the process transfer function at the critical pulsation is deduced

$$G_{proc}(j\omega_c) = -\frac{\pi a_y}{4a_u} \tag{4.16}$$

The ultimate gain K_{cu} of the proportional controller is thus the reciprocal of the modulus of $G_{bo}(j\omega_c)$

$$K_{cu} = \frac{4a_u}{\pi a_y} \tag{4.17}$$

The condition of oscillation is that the curves $G_{proc}(j\omega)$ and $-1/G_{rel}$ intersect in the complex plane, which corresponds to a Nyquist diagram.

From the knowledge of the ultimate period $T_c = T_u$ and the ultimate gain K_{cu}, the initial tuning of the controller can be performed following Ziegler-Nichols tuning recommendations (Table 5.2).

This relay technique has been improved and extended for processes with delay (Scali et al., 1999) and multivariable processes (Semino and Scali, 1998).

Example 4.1: **Use of the Relay Oscillation Method**

A third-order system presents an oscillation-critical frequency. Consider the same system as in Sect. 3.3.3 which contains an actuator represented by a first-order $G_a = 1/(s+1)$, a process represented by a second-order $G_p = 1/(s^2 + 10s + 24)$ and a transmitter that has a pure gain equal to 8. In the absence of a relay, the controller is set in proportional mode with a gain equal to 30. This gain does not give a zero steady-state deviation, but with a slightly higher gain, the process becomes unstable. Then, a proportional controller is taken with unity gain (if the gain was left equal to 30, the result would be very close) and the system is subjected to a step unit since $t = 0$ and the relay placed before the actuator is operated at the same time. The relay possesses the following characteristics: its state changes from closed to open for an input value taken as 0.001 (a small value corresponding to a zero error signal) and from open to closed for the same value (taking 0.001 or -0.001 as in the previous theory would have very little influence), while the value corresponding to the actuator input around the steady state is about 22. When the relay state is

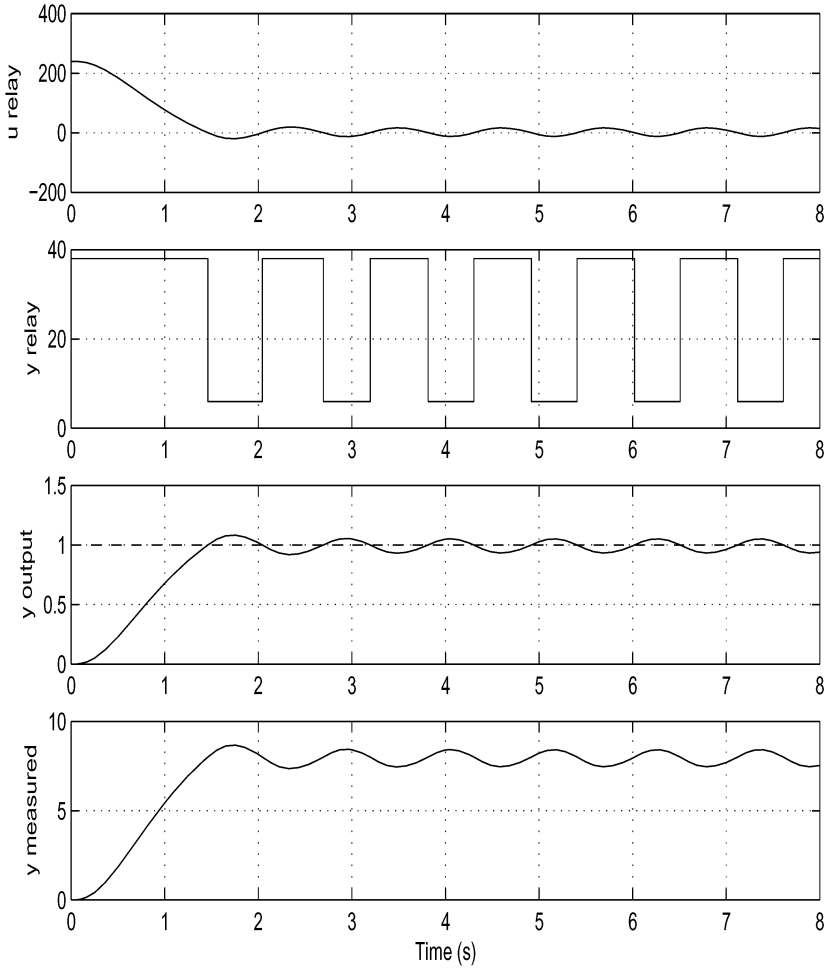


Figure 4.7: Responses obtained in the case of the use of a relay to make a third-order system oscillate. Top curve: relay input. Middle top curve: relay output. Middle bottom curve: process output y with set point. Bottom curve: measured output y_m

closed, its output is open to 38; when its state is open, its output is equal to 6, so that the amplitude of the rectangular relay wave is equal to 16. It is thus observed that it is possible to make the system oscillate around a given state, which is indeed its set point (Fig. 4.7) with a frequency $\omega_c = 2\pi/T_u$, in the present case: $\omega_c = 2\pi/1.096 \approx 5.73$ rad/time unit. The frequency thus found is near the critical frequency found in Sect. 3.3.3. Because of the nonlinearities introduced by the relay and the slightly different proportional gain, it is not exactly the same.

The amplitude of the relay output is $a_u = 16$ while the amplitude of the measured output is $a_y \approx 0.473$. Thus, the ultimate gain of the proportional controller given by Eq. (4.17) is $K_{cu} \approx 43.07$, close to the exact value found in Sect. 3.3.3. Knowing the ultimate period T_u and the ultimate gain K_{cu}, the initial tuning of the controller can be easily performed following Ziegler-Nichols tuning recommendations (Table 5.2).

The state around which the oscillations occur corresponds to the steady state provided that a sufficiently high proportional gain is chosen to avoid the steady-state offset.

Theoretically, only systems of order larger than 2 or having a time delay can possess a phase angle which can become lower than $-180°$ above some frequency. In practice, nearly all systems present a time delay, which even can be low, and thus can oscillate in this manner.

4.5.4 Process Reaction Curve Method

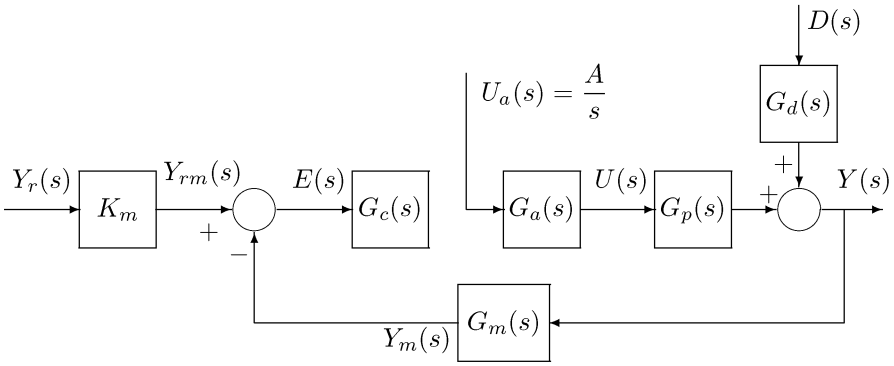


Figure 4.8: Open control loop for Cohen and Coon method

To get the process reaction curve, Cohen and Coon (1953) recommend opening the control loop (Fig. 4.8) between the controller and the actuator and imposing a step on the actuator input u_a. Between the Laplace transform of the measured output Y_m and that of the actuator input U_a, the transfer function corresponding to the process reaction curve is

$$G_{prc} = \frac{Y_m}{U_a} = G_a\, G_p\, G_m \tag{4.18}$$

Thus, the dynamics of the measured output depends not only on the process, but also on the actuator and on the measurement device. These three elements, process, actuator and sensor, constitute the physical environment which cannot be dissociated to get a measurement value with respect to a given actuator position.

The measured responses y_m frequently have a sigmoidal allure observed in Fig. 4.9. This curve is approximated by a curve corresponding to a first-order system with delay transfer function (Fig. 4.9)

$$G_{prc} = \frac{Y_m}{U_a} \approx \frac{K\ \exp(-t_d\, s)}{\tau s + 1} \tag{4.19}$$

Thus, three parameters must be estimated from the experimental curve:

- $K = B/A$ with B as the steady-state output, and A as the amplitude of the input step
- $\tau = B/S$ with S as the slope of the sigmoidal response at the inflection point (note that this determination of S is imprecise and results in an approximate value of τ).
- t_d time gone before the system response

One thus obtains the curve approaching the sigmoid. This method is also called Broida's method. If τ is difficult to estimate practically, it can be well approximated by the residence time τ_{res}

$$\tau_{res} = \frac{\displaystyle\int_0^\infty t g(t) dt}{\displaystyle\int_0^\infty g(t) dt} = -\frac{G'(0)}{G(0)} \tag{4.20}$$

where $g(t)$ is the impulse response of a stable process having the transfer function $G(s)$.

It would be possible to choose an approximate model of a delayed second-order transfer function, which would provide a better approximation

$$G_{prc} = \frac{Y_m}{U_a} \approx \frac{K\ \exp(-t_d\, s)}{(\tau_1 s + 1)(\tau_2 s + 1)} \qquad \text{with:} \quad \tau_1 > \tau_2 \tag{4.21}$$

The gain K and the delay t_d are obtained in the same manner as previously discussed. The time constants τ_1 and τ_2 can be identified by a least-squares method or approached by the moments method or Harriott's method.
Then the parameter values of the approximation transfer function must be used to deduce the tunings of different controllers. Assuming a delayed first-order transfer function, to get the below-mentioned values, Cohen and Coon (1953) have used the following criteria:

- One quarter decay ratio.
- Minimum steady-state offset.
- Minimum integral of square error (ISE).

They obtained the following recommended tunings:
a) Proportional controller:

$$K_c = \frac{1}{K}\ \frac{\tau}{t_d}\ \left(1 + \frac{t_d}{3\,\tau}\right) \tag{4.22}$$

b) PI controller:

$$K_c = \frac{1}{K}\ \frac{\tau}{t_d}\ \left(0.9 + \frac{t_d}{12\,\tau}\right) \tag{4.23}$$

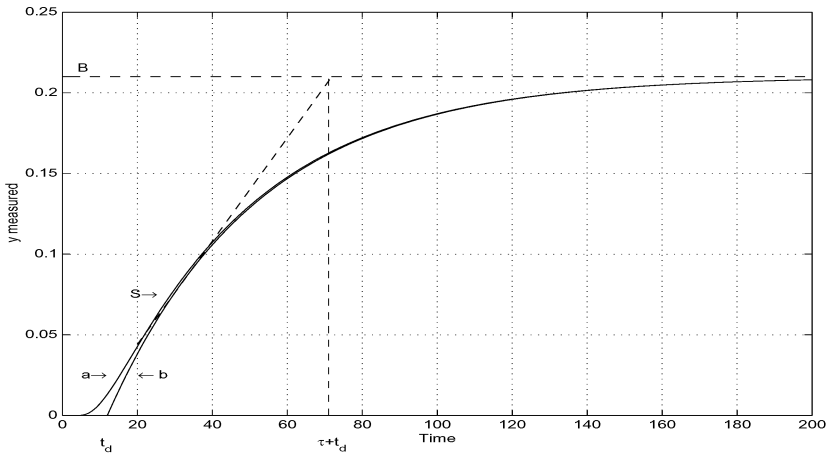


Figure 4.9: Parameter determination according to the process reaction curve method recommended by Cohen and Coon (sigmoidal response: curve a, approximation by a first-order transfer function with time delay: curve b)

$$\tau_I = t_d \frac{30 + 3\,t_d/\tau}{9 + 20\,t_d/\tau} \tag{4.24}$$

c) PID controller:

$$K_c = \frac{1}{K}\,\frac{\tau}{t_d}\,\left(\frac{4}{3} + \frac{t_d}{4\,\tau}\right) \tag{4.25}$$

$$\tau_I = t_d \frac{32 + 6\,t_d/\tau}{13 + 8\,t_d/\tau} \tag{4.26}$$

$$\tau_D = t_d \frac{4}{11 + 2\,t_d/\tau} \tag{4.27}$$

These values must not be considered as final values, but as initial values of controller tuning, in particular when the response given by the first-order system with time delay goes far from the open-loop system response curve. According to De Larminat (1993), a PID controller gives excellent results when the ratio τ/t_d is large (larger than 5 or 10). If $t_d > 0.5\tau$, i.e. the delay is relatively large compared to the process-dominant time constant, it must be considered that a simple PID controller is not appropriate.

The following generalities valid for other tunings and relative to P, PI and PID controllers can be observed:

- The gain of the PI controller is smaller than that of the P controller (because of the tendency of integral action to destabilize the process).
- On the other hand, the gain of the PID controller is larger than those of the P and PI controllers, because the derivative action stabilizes the process.

4.5.5 Tuning rule of Tavakoli and Fleming for PI controllers

The tuning of PI controllers for first order systems with delay proposed by (Tavakoli and Fleming, 2003) guarantee a minimum gain margin of 6dB and a minimum phase margin of 60° even when the delay t_d is large with respect to the time constant τ of the system. The rules are

$$\begin{aligned} K\,K_c &= 0.4849\,\frac{\tau}{t_d} + 0.3047 \\ \frac{\tau_I}{\tau} &= 0.4262\,\frac{\tau}{t_d} + 0.9581 \end{aligned} \tag{4.28}$$

where K is the gain of the open-loop system. This method gives results close to that proposed by (Hang et al., 1991).

4.5.6 Robust tuning rule for PID controllers

(Aström and Hagglund, 2004) have conducted many studies on PID control. Recently, they proposed approximate M-Constrained integral gain optimization (AMIGO) to increase the robustness expressed through the parameter M. Their PID controller presents a formula slightly different from the classical PID as

$$r(t) = K_c \left[(by_r(t) - y_f(t)) + \frac{1}{\tau_I} \int_0^t (y_r(x) - y_f(x))dx + \frac{1}{\tau_D} \left(c\frac{dy_r(t)}{dt} - \frac{dy_f(t)}{dt} \right) \right] \tag{4.29}$$

where $r(t)$ is the output signal from the controller, y_r the set point, y_f the filtered output given by the filter transfer function

$$G_f(s) = \frac{Y_f(s)}{Y(s)} = \frac{1}{(T_f s + 1)^2} \qquad \text{or:} \quad G_f(s) = \frac{1}{T_f s + 1} \tag{4.30}$$

The second-order filter is chosen to increase the filtering action. Parameters b and c in equation (4.29) are called set point weightings which influence the response to set point changes. The controller thus designed with set point weighting has two degrees of freedom: set point and disturbance responses are both taken into account. For a process described by a first-order transfer function with delay such as (4.19), the relative dead time is defined as

$$\tau_r = \frac{t_d}{t_d + \tau} \tag{4.31}$$

and the AMIGO tuning rules for a PID controller are

$$K_c = \frac{1}{K} \left(0.2 + 0.45\frac{\tau}{t_d} \right) \tag{4.32}$$

$$\tau_I = t_d\,\frac{0.4\,t_d + 0.8\,\tau}{t_d + 0.1\,\tau} \tag{4.33}$$

$$\tau_D = \frac{0.5\, t_d\, \tau}{0.3\, t_d + \tau} \tag{4.34}$$

When $\tau_r > 0.3$, these rules give good results. For lower values, the gain K_c must be increased, τ_I decreased and τ_D increased. However, the tuning is robust for all processes. A conservative choice of parameter b is

$$b = \begin{cases} 0 & \text{for } \tau \leq 0.5 \\ 1 & \text{for } \tau > 0.5 \end{cases} \tag{4.35}$$

while in general $c = 0$ except for smooth set point changes.

For an integrating process represented by the transfer function

$$G_{prc}(s) = \frac{K\, \exp(-t_d s)}{s} \tag{4.36}$$

the AMIGO tuning rules for a PID controller become

$$K_c = \frac{0.45}{K} \quad ; \quad \tau_I = 8\, t_d \quad ; \quad \tau_D = 0.5\, t_d \tag{4.37}$$

4.6 PID Improvement

As opposed to the ideal PID, a real PID, as already discussed in Sect. 2.3.4, with transfer function $G_c(s)$

$$G_c(s) = K_c \left(\frac{\tau_I s + 1}{\tau_I s}\right) \left(\frac{\tau_D s + 1}{\beta\, \tau_D s + 1}\right) \tag{4.38}$$

or slightly modified transfer function $G_c(s)$

$$G_c(s) = K_c \left(1 + \frac{1}{\tau_I s} + \frac{\tau_D s}{\frac{\tau_D}{N} s + 1}\right) \tag{4.39}$$

should be considered as the minimum PID controller implementable on a process.

With respect to these versions, variants of PID controllers have been proposed in order to solve some problems which can occur in the use of classical PID.

4.6.1 PID Controller with Derivative Action on the Measured Output

It is often preferable (Aström and Hägglund, 1988; Wolovich, 1994) to operate the PID controller by making the derivative action act no more on the error coming from the comparator but on the measured output (Fig. 4.10), under the theoretical form

$$u_a(t) = K_c \left(e(t) + \frac{1}{\tau_I} \int_0^t e(x)\, dx - \tau_D \frac{dy_m}{dt}\right) \tag{4.40}$$

or practically

$$U_a(s) = K_c \left(1 + \frac{1}{\tau_I s}\right) E(s) - \left(\frac{K_c\, \tau_D\, s}{\frac{\tau_D}{N}\, s + 1}\right) Y_m(s) \tag{4.41}$$

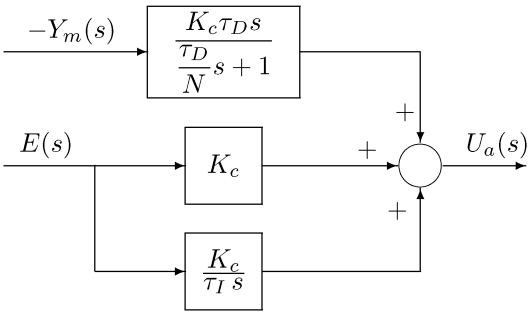


Figure 4.10: Block scheme of the PID controller with derivative action on the measured output

This manner of taking into account the derivative action enables us to avoid steep changes of the controller output due to the error signal variation during set point changes. This controller will be called PID with derivative action on the measurement. The theoretical derivative action is filtered by a first-order system so that the actual derivative action acts especially on the low-frequency signals; the high-frequency measurement is, at the most, amplified by coefficient N.

4.6.2 Use of a Reference Trajectory

A violent set point change, for example step-like, induces fast and important variations of control variable u, which can be mechanically harmful to the actuator or which can lead the control to saturation. In order to minimize the brutal effect of set point change, it is possible to filter the set point by a unit gain first-order or overdamped second-order filter G_{ref} (Fig. 4.11) so that the output y will be compared to a smoothed set point y_{ref}, called a reference trajectory. In Fig. 4.12, it is easy to note that:

- In the case of the use of this second-order reference trajectory, the output reacts more slowly, but does not present any overshoot anymore.
- The control variable varies much less brutally and in a narrower range.

4.6.3 Discretized PID Controller

In a system controlled by a computer, measurements are made at discrete times t_k such that two successive instants are separated by a sampling period T_s. Without entering into the details of the choice of the sampling period (Table 9.3), it is useful here to mention the incremental form of a PID controller

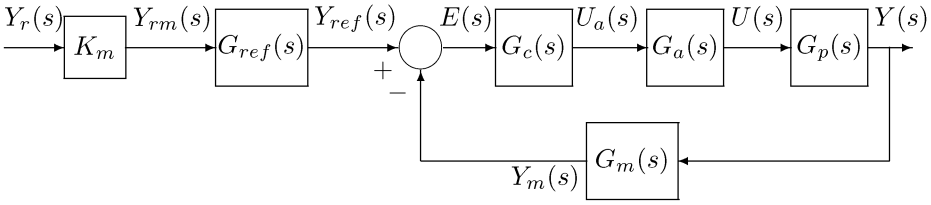


Figure 4.11: Control with reference trajectory

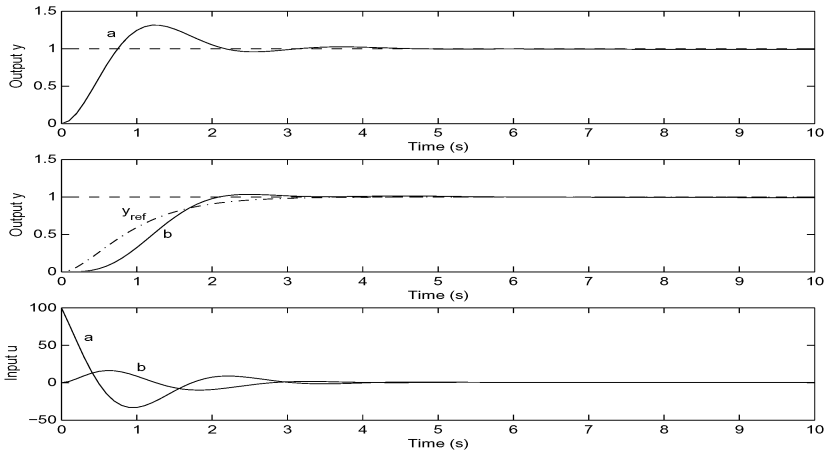


Figure 4.12: Upper figure: response to a set point unit step without reference trajectory (curve a), middle figure: Reference trajectory and response to a set point unit step with reference trajectory (curve b), lower figure: corresponding controls. Process ($K_p = 2, \tau = 5, \zeta = 0.5$) ; real PID controller ($K_c = 10, \tau_I = 5, \tau_D = 4, \beta = 0.1$); second-order reference trajectory ($K = 1, \tau = 0.5\,, \zeta = 1$)

(Aström and Hägglund, 1988). From a given continuous form, several forms can be obtained according to the type of discretization (backward difference, forward difference, etc.). The Tustin algorithm, which gives the nearest form to the continuous controller is used in the form described below (Aström and Hägglund, 1988). The control is expressed by its variation between two successive instants

$$\Delta u(t_k) = u(t_k) - u(t_{k-1}) = \Delta P(t_k) + \Delta I(t_k) + \Delta D(t_k) \tag{4.42}$$

where P, I, D represent the following proportional, integral and derivative contributions

$$\Delta P(t_k) = P(t_k) - P(t_{k-1}) = K_c\ (y_{rd}(t_k) - y_m(t_k) - y_{rd}(t_{k-1}) + y_m(t_{k-1})) \tag{4.43}$$

The variable y_{rd} (reference) is equal to the set point y_r if no reference trajectory (set point filtering) is used; otherwise, it is equal to y_{ref}, filtered set point or

reference trajectory. Aström and Hägglund (1988) propose to take y_{rd} as

$$y_{rd} = b\, y_r \quad ; \quad 0 \leq b \leq 1 \tag{4.44}$$

where b is simply a constant. The choice $b = 0$ gives a sluggish response without overshoot, $b = 1$ (no filtering) gives a fast response with a possible overshoot. The integral term is equal to

$$\Delta I(t_k) = I(t_k) - I(t_{k-1}) = \frac{K_c\, T_s}{\tau_I}\,(y_r(t_{k-1}) - y_m(t_{k-1})) \tag{4.45}$$

Note that it would have been possible to replace y_r by y_{rd}.
The derivative term is equal to

$$\Delta D(t_k) = D(t_k) - D(t_{k-1}) = \frac{b}{1-a}\,(y_m(t_k) - 2\, y_m(t_{k-1}) + y_m(t_{k-2})) \tag{4.46}$$

with the coefficients a and b equal to

$$a = \frac{2\,\tau_D - T_s\, N}{2\,\tau_D + T_s\, N} \quad ; \quad b = -\frac{2\, K_c\, N\, \tau_D}{2\,\tau_D + T_s\, N} \tag{4.47}$$

such taht

$$D(t_k) = a\, D(t_{k-1}) + b\,(y(t_k) - y(t_{k-1})) \tag{4.48}$$

It is necessary to choose $\tau_D > N\, T_s/2$ (otherwise $a < 0$). The prescribed form amounts to filtering the ideal derivative action by a first-order system with time constant τ_D/N. If N is chosen to be large, the high-frequency measurement noise is amplified.

4.6.4 Anti-Windup Controller

For the purpose of anti-windup, a controller based on a real PID controller with derivative action on the measured output is used.

The actuator presents physical limitations (a valve position is situated between full opening and full closure), so that when the control variable imposes the actuator position at its limits, the feedback loop becomes inefficient as the control is saturated and cannot vary anymore. In this case, the error becomes, in general, important and the integral term even more so. The integrator is subjected to windup; this is also called integral saturation. The solution is to stop the action of this integrator as long as saturation lasts. To avoid the undesirable windup phenomenon, it is possible to add an additional feedback loop by using the difference e_a between the controller output and the actuator model output (Fig. 4.13) (Aström and Hägglund, 1988; Hanus, 1990, 1996; Hanus and Peng, 1992; Hanus et al., 1987). In the absence of actuator saturation, this difference e_a is zero. When the actuator saturates, the proposed anti-windup system acts to try to bring back this difference e_a towards zero. The higher the loop gain (the time constant is low), the faster this system reacts.
The anti-windup system has been tested in the following configuration:

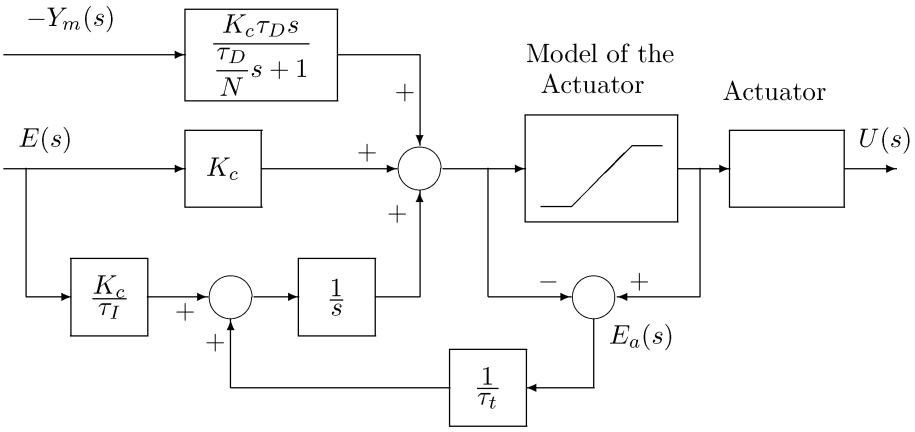


Figure 4.13: PID controller with anti-windup device

- The process is an underdamped second-order system presenting the following characteristics: $K_p = 2$, $\tau_p = 5$, $\zeta = 0.5$.
- The actuator saturates when the input is out of interval $[-1, 1]$.
- The used controller is a real PID with derivative action on the measured output: $K_c = 30$, $\tau_I = 2$, $\tau_D = 2$, $N = 5$.
- The system is subjected at time $t = 1$ to a set point unit step.

In the absence of an anti-windup system (it suffices to choose gain $1/\tau_t$ zero), the actuator saturates during long periods, both on the positive and negative sides, which provokes a large overshoot of the output with respect to the set point (Fig. 4.14). When the anti-windup system is installed with gain $1/\tau_t = 10$, the actuator still saturates at the beginning, but during a much shorter period, and the output joins the set point back smoothly with very little overshoot (Fig. 4.14).

It can be noticed that, with an integral gain very slightly higher (with $\tau_I = 1,5$), when the anti-windup system does not work, the system is near instability. If the anti-windup system is operated, this problem disappears. Similarly, by imposing still more severe constraints on the actuator, it becomes very difficult, or even impossible, to function without anti-windup system, while with the latter, the process mastering is realized without difficulty.

4.6.5 PID Control by On−Off Action

In many processes, it may happen that the actuator offers only two possibilities: on or off. Such a case occurred in a lab crystallizer, where heating is achieved by electrical resistance and the only possibility is to switch the electric current on or off. If a simple controller is used, it will result in important oscillations around the set point, even more so as the process is highly nonlinear because heating is performed by radiative heat transfer while cooling is, in general, performed by a coil or a jacket or even natural convection for an oven for

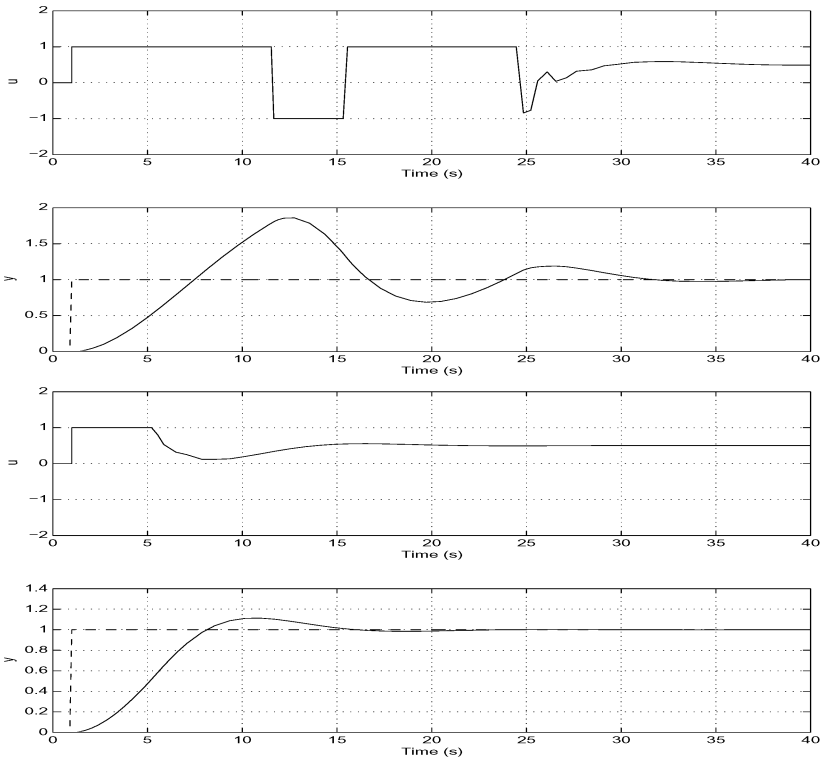


Figure 4.14: Top figure: input u and output y in response to a set point unit step without anti-windup system ($1/\tau_t = 0$), lower figure: input u and output y in response to a set point unit step with anti-windup system ($1/\tau_t = 10$). Process ($K_p = 2, \tau = 5, \zeta = 0.5$); real PID controller ($K_c = 30, \tau_I = 2, \tau_D = 2, N = 5$)

example.

The following strategy allows us to deal with such a case. First, a base period t_b is chosen and this period is divided into n sub-intervals $t_c = t_b/n$, which will be the control period: the actuator will be allowed to change at each new sub-interval.

An integer u_h takes the value 1 if heating is desired, 0 if no heating.

An internal counter *inter* is added, which counts from 0 to n. When the counter reaches n, it is reinitialized to 0. The counter advances by one unit at every t_c. A digital clock gives the time t_d. The counter is equal to

$$inter = \text{int}(t_d/t_c) - \text{int}(t_d/(t_c\, n))\, n \tag{4.49}$$

where "int" indicates the integer part of a variable.

A positive maximum error e_{max} is given by the user. It corresponds to a very large error due to underheating at the beginning of operation for which

we can accept that the heating is on during all t_b. A positive minimum error e_{min} is also given. It corresponds to the absolute error due to such overheating that the heating must be switched off.

The error $e = T_r - T$ between the reference temperature and the crystallizer temperature is normalized by the maximum error so that its absolute value is between 0 and 1

$$e_n = \frac{e}{e_{max}} \tag{4.50}$$

A fraction of heating is given by

$$i_h = \text{int}(e_n\, n) \tag{4.51}$$

which is the number of sub-intervals t_c with respect to t_b during which it is necessary to heat.

Then, a rule providing the value of h must be chosen

$$\begin{array}{ll} \text{if } e > e_{max} & \text{then } u_h = 1 \\ \text{else if } e < -e_{min} & \text{then } u_h = 0 \\ \text{else if } inter < i_h & \text{then } u_h = 1 \\ \text{else if } inter > i_h & \text{then } u_h = 0 \end{array} \tag{4.52}$$

That set of rules can be implemented between a normal PID controller and the actuator. In our case, a discrete PI controller was used and the error $e(t)$ used in the previous algorithm was replaced by the output of the controller. The gain of the proportional controller should be near 1.

Example 4.2: **On-Off Temperature Control of a Batch Crystallizer**

Before implementation in the real process, a simulation model of the crystallizer

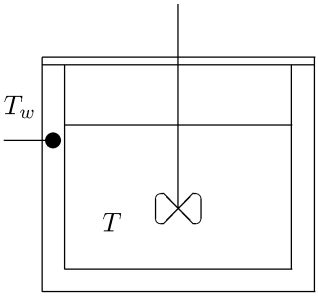


Figure 4.15: Batch crystallizer

was designed. The state-space model of the batch crystallizer (Fig. 4.15) used

for simulation is

$$\begin{aligned} \frac{dT}{dt} &= \frac{UA_{int}\,(T_w - T)}{V_l\,\rho_l\,C_{p,l}} - \frac{UA_{ext}\,(T - T_{amb})}{V_l\,\rho_l\,C_{p,l}} \\ \frac{dT_w}{dt} &= \frac{P_{eff}}{V_m\,\rho_m\,C_{p,m}} - \frac{UA_{int}\,(T_w - T)}{V_m\,\rho_m\,C_{p,m}} \\ \frac{dP_{eff}}{dt} &= \frac{u_h\,P - P_{eff}}{\tau_r} \end{aligned} \tag{4.53}$$

with the parameter values given in Table 4.4.

Table 4.4: Main parameters and initial variables of the crystallizer

Reactor volume	$V = 4.5 \times 10^{-3}$ m^3
Density of reactor contents	$\rho = 1000$ kg.m^{-3}
Heat capacity of reactor contents	$C_p = 4185$ J.kg^{-1}.K^{-1}
Internal global heat transfer coefficient	$UA_{int} = 14$ W.m^{-2}.K^{-1}
External global heat transfer coefficient	$UA_{ext} = 0.7$ W.m^{-2}.K^{-1}
Volume of wall metal	$V_m = 0.9 \times 10^{-3}$ m^3
Density of wall metal	$\rho_m = 8055$ kg.m^{-3}
Heat capacity of wall metal	$C_{p,m} = 490$ J.kg^{-1}.K^{-1}
Heat power of the electrical resistance	$P = 1750$ W
Time constant of the electrical resistance	$\tau_r = 5$ s
Initial temperature in the crystallizer	$T = 290.15$ K
Initial temperature in the crystallizer wall	$T_w = 290.15$ K
Initial effective heat power	$P_{eff} = 0$ W

The parameters were chosen as $n = 30$, $t_c = 1$, and for the PI controller: $K_c = 0.5$, $\tau_I = 10000$. Figure 4.16 shows that the temperature regulation is well ensured.

4.6.6 pH Control

The pH control in neutralization processes poses difficult problems because of the extreme nonlinearity of the pH curve during neutralization (Fig. 4.18). Shinskey (1988) devotes an important part of his book to pH control and proposes several solutions to practically remedy those problems. Lee et al. (2001) conduct an important review of the various propositions for pH control. Many different approaches for pH control have been proposed in the literature. Even in an early paper, McAvoy et al. (1972) used the balances and equilibrium equations. When several weak and strong acids or bases are part of a multicomponent medium, knowledge of the chemical equilibria can be intricate. The concept of invariant reaction was introduced by Gustafsson and Waller (1983) and then extended (Gustafsson et al., 1995). Other approaches based

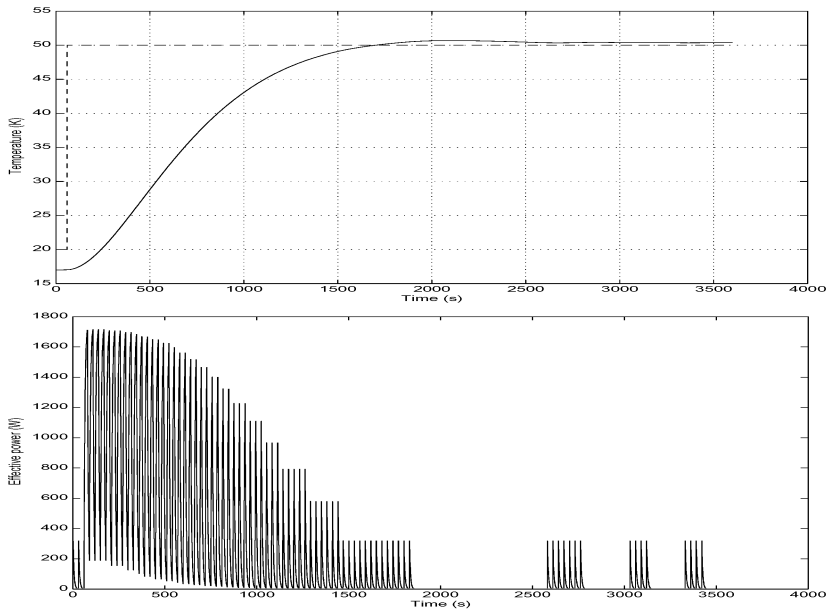


Figure 4.16: Heating of the crystallizer. Top figure: set point and temperature in the crystallizer. Lower figure: effective power during heating of the crystallizer

on identification have been proposed, and adaptive control is frequently cited (Gustafsson and Waller, 1992; Lee et al., 1993, 1994; Sung et al., 1998). Nonlinear control is proposed by Wright and Kravaris (1991). Often, identification reactors (Gupta and Coughanowr, 1978) or in-line mixers are proposed (Sung et al., 1995), eventually coupled with feedforward control (Lee and Choi, 2000). Choi et al. (1995) propose a process simulator for pH control studies.

As a simple example, consider a fed-batch chemical reactor (Fig. 4.17). At the initial time, this reactor contains $V_0 = 10$ cm^3 of 0.1N hydrochloric acid HCl. The neutralization is performed by the use of 0.1N caustic soda NaOH with an inlet flow rate F_{in} (0.05 cm^3/s). Though these are respectively a strong acid and base, their concentration is rather low. The chemical reaction is

$$\mathrm{H^+\,Cl^-} \;+\; \mathrm{Na^+\,OH^-} \rightarrow \mathrm{Na^+\,Cl^-} \;+\; \mathrm{H_2\,O}$$

Assuming an inlet flow rate F_{in} (cm^3/s), the different balances read

$$\begin{aligned} \frac{dV}{dt} &= F_{in} \\ \frac{d[\mathrm{Na^+}]}{dt} &= \frac{F_{in}}{V}\left([\mathrm{OH^-}]_{in} - \mathrm{Na^+}]\right) \\ \frac{d[\mathrm{Cl^-}]}{dt} &= -\frac{F_{in}}{V}\,[\mathrm{Cl^-}] \end{aligned} \tag{4.54}$$

to which we must add the equilibrium constant $K_{eq} = 10^{-14}$ (at 25 °C) for

water dissociation

$$[H^+]\,[OH^-] = K_{eq} \tag{4.55}$$

and the electro-neutrality equation

$$[Na^+] + [H^+] = [Cl^-] + [OH^-] \tag{4.56}$$

The H^+ concentration results from this equation, from which the pH is deduced

$$pH = -\log_{10}([H^+]) \tag{4.57}$$

The solution of the previous equations gives Fig. 4.18 which is typical of the neutralization of a strong acid by a strong base where the desired pH is the neutral pH equal to 7. The control question is the following: for any type of reactor, fed-batch or continuous, perfectly stirred or plug flow, how is it possible to maintain the pH in the neighbourhood of a given value?

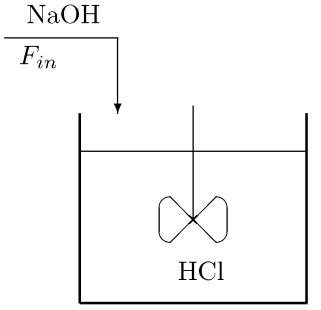


Figure 4.17: Fed-batch neutralization reactor in initial conditions

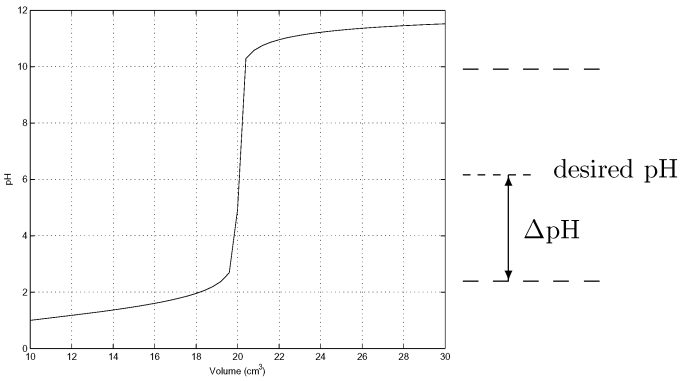


Figure 4.18: Typical pH variation during neutralization

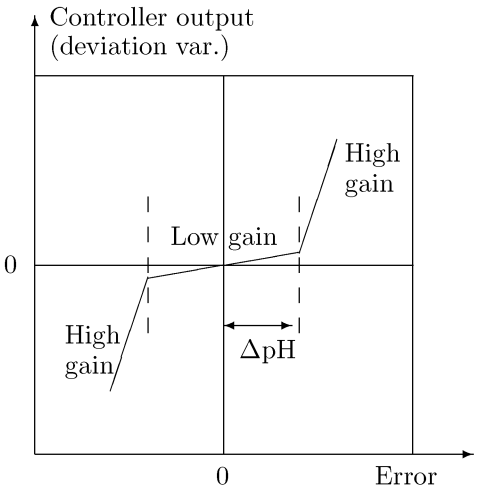


Figure 4.19: Zones of different controller gain according to pH

Around the neutralization point, for very small amounts of added reactant, the pH changes very rapidly, while far from the neutralization point, large amounts of reactant provoke a small change. Under these conditions, a classical controller with a fixed gain is unable to operate correctly. Moreover, the shape of the neutralization curve depends on the chemical components and it is recommended to have an idea by titration of the chemical species involved. Generally speaking, the controller can be composed into three parts (Fig. 4.19), one in a zone well below the desired pH, another one well above the desired pH where a high gain can be used, and a third zone around the desired pH where a low gain is necessary. If a linear controller with a single gain is chosen, there exist two possibilities: either the controller gain is very low and the corresponding flow of neutralizing agent is also low, resulting in very long time responses far from the desired pH, or the controller gain is high and the flow of neutralizing agent is large, inducing rapid variations around the pH set point, which results in limit cycles and the pH oscillating from low to high values. The proposed nonlinear controller with three zones does not present that drawback. The ratio between the high and the low gains can be as large as 1000; it depends on the type of chemical acids and bases that are present. Thus, the controller can only be tuned after titration of the different media and its tuning must be modified if the nature of the influent changes. For those reasons, adaptive and feedforward controllers can be proposed for pH control. Shinskey (1988) also proposes to use two equal-percentage valves, one for a large flow rate, the second for a flow rate slightly over the minimum capability of the first one. Thus, by operating only one of the two valves at any time, the very different conditions of flow can be achieved with negligible interruption during the passage from one valve to the other.

Example 4.3: **Control of a Neutralization Reactor**

Consider a continuous neutralization reactor (Fig. 4.20) which has two efflu-

ents, one acid represents the stream to be neutralized and is a disturbance, and one base, which is the neutralization solution and the manipulated input. As a simple example, the acid is assumed to be a strong acid such as HCl and the base is a strong base such as NaOH.

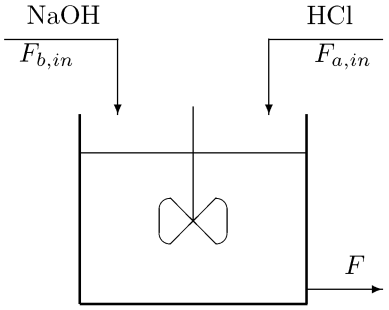


Figure 4.20: Continuous neutralization reactor

The acid inlet flow rate is denoted by $F_{a,in}$ (m^3/s) and the base inlet flow rate is $F_{b,in}$ (m^3/s). The level control is assumed to be perfect. With respect to the previous fed-batch reactor, the different balances are slightly modified

$$\begin{aligned} F_{a,in} + F_{b,in} &= F_{out} \\ \frac{d[\mathrm{Na}^+]}{dt} &= \frac{1}{V}\left(F_{b,in}\,[\mathrm{OH}^-]_{in} - F\,[\mathrm{Na}^+]\right) \\ \frac{d[\mathrm{Cl}^-]}{dt} &= \frac{1}{V}\left(F_{a,in}\,[\mathrm{Cl}^-] - F\,[\mathrm{Cl}^-]\right) \end{aligned} \tag{4.58}$$

The equilibrium equation for water dissociation and the electro-neutrality equation are not modified.

A variable-gain PI controller has been used for controlling this reactor according to the conditions given in Table 4.5.

Table 4.5: Main parameters and initial variables of the neutralization reactor

Acid inlet flow rate	$F_{a,in} = 20$ l/s for t in [0,400]s then 40 l/s for t in [400,600]s
Acid inlet concentration	$C_{a,in} = 0.2$ mol/l for t in [0,200]s then 0.4 mol/ for t in [200,600]s
Base inlet concentration	$C_{b,in} = 1$ mol/l
Reactor volume	$V = 10^4$ l
Initial pH	$\mathrm{pH}(t = 0) = 1$
High proportional gain	$K_c = 500$
Low proportional gain	$K_c = 50$
Integral time constant	$\tau_I = 100$ s
Maximum base flow rate	$max(F_{b,in}) = 100$ l/s

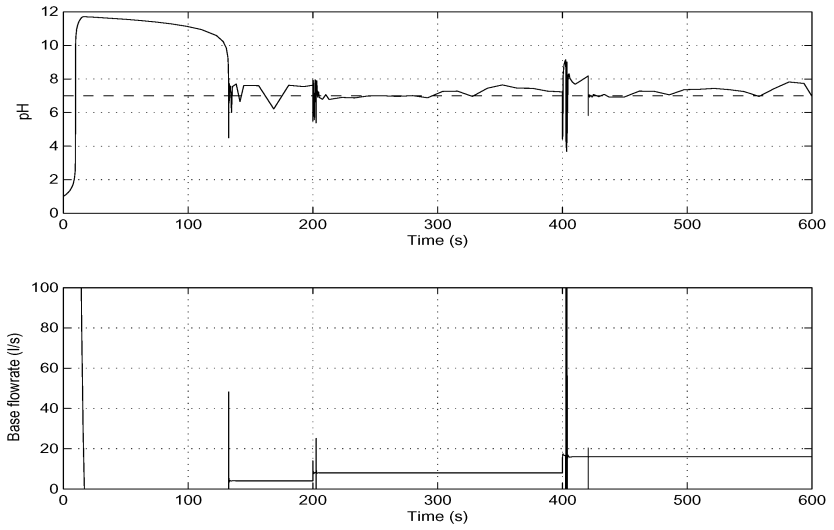


Figure 4.21: pH control of a neutralization reactor. Top: pH controlled. Bottom: inlet flow rate of base

Initially, the continuous neutralization reactor is far from the steady state. From Fig. 4.21, it appears that at the beginning there is a saturation of the manipulated base flow rate. Then a stabilization with oscillations occurs around the desired neutral pH. The acid concentration disturbance at 200 s and the acid flow rate disturbance at 400 s are both correctly rejected.

4.7 Direct Synthesis Method

Considering the closed-loop block scheme (Fig. 2.19), one obtains the relation

$$Y(s) = \frac{G_p(s)\,G_a(s)\,G_c(s)\,K_m}{1 + G_p(s)\,G_a(s)\,G_c(s)\,G_m(s)}\,Y_r(s) + \frac{G_d(s)}{1 + G_p(s)\,G_a(s)\,G_c(s)\,G_m(s)}\,D(s) \tag{4.59}$$

hence the closed-loop transfer function corresponding to the set point variations

$$\frac{Y}{Y_r} = \frac{G_p\,G_a\,G_c\,K_m}{1 + G_p\,G_a\,G_c\,G_m} \tag{4.60}$$

Set

$$G = G_a G_p \tag{4.61}$$

One deduces

$$\frac{Y}{Y_r} = \frac{G G_c K_m}{1 + G G_c G_m} \tag{4.62}$$

hence the "theoretical" controller transfer function

$$G_c = \frac{1}{G} \frac{\frac{Y}{Y_r}}{K_m - G_m \frac{Y}{Y_r}} \tag{4.63}$$

In fact, the process transfer function G_p, thus globally G, is not perfectly known and the ratio Y/Y_r is unknown because the controller has not yet been designed. This imposes replacement of the transfer function G by the supposed or estimated process model $\tilde{G}$ and the ratio Y/Y_r by a desired ratio $(Y/Y_r)_d$.

In this case, one gets

$$G_c = \frac{1}{\tilde{G}} \frac{(\frac{Y}{Y_r})_d}{K_m - G_m (\frac{Y}{Y_r})_d} \tag{4.64}$$

It is remarkable that the process model appears in G_c through its inverse; this feature is one of the characteristics of model-based control techniques. This will induce a given number of problems concerning poles and zeros: the controller poles will be the process zeros and vice versa. It will be necessary to check whether this controller is stable and physically realizable.

The direct synthesis technique is essentially of theoretical interest and should be carefully used for processes having unstable poles or zeros. Internal model control, which follows in Sect. 4.8, presents some common features and should be preferred.

4.8 Internal Model Control

The controller design method called internal model control has been in particular developed and popularized by Garcia and Morari (1982). At this point, we emphasize that it differs from the principle of internal model by Francis and Wonham (1976) explained in Sect. 5.9 and it is useful to compare both ideas. Internal model control is based on the knowledge of a supposed model of the process. The model uncertainty is directly taken into account. It is possible to compensate the performance of the control system by its robustness to process modifications or modelling errors. This control is recommended in particular when the ratio τ/t_d from Cohen and Coon (1953) is small (lower than 5).

In the block diagram (Fig. 4.23), just as for direct synthesis, the process transfer function G represents in fact the process itself and its instrumentation (actuator and measurement). The process model $\tilde{G}$ obtained by identification and the controller output U_a allow us to calculate $\tilde{Y}$. In general, the real process G and its identified model $\tilde{G}$ differ, moreover disturbances D influencing Y are unknown, thus $\tilde{Y}$ is different from Y.

In a classical feedback control scheme (Fig. 4.22), the output is related to the set point and to the disturbance by

$$Y = \frac{G\,G_c\,Y_r + D}{1 + G\,G_c} \tag{4.65}$$

In internal model control (Fig. 4.23), denoting the calculated controller by G_c^*, the following relations are obtained

$$Y = G\,U_a + D \tag{4.66}$$

$$\tilde{Y} = \tilde{G}\,U_a \tag{4.67}$$

$$U_a = G_c^*\,E = G_c^*\,(Y_r - Y + \tilde{Y}) = G_c^*\,(Y_r - Y + \tilde{G}\,U_a) \tag{4.68}$$

hence

$$(1 - \tilde{G}\,G_c^*)\,U_a = G_c^*(Y_r - Y) \tag{4.69}$$

Two characteristics are desired by the user:

- A perfect set point tracking, thus $y = y_r$, when no account is made of disturbances.
- A maximum rejection of disturbances, thus a minimum influence of d, when the mode is regulation.

In these conditions, to get a perfect set point tracking which is equivalent to: $Y(s) = Y_r(s)$ when $D = 0$, it suffices that

$$(1 - \tilde{G}\,G_c^*) = 0 \Longleftrightarrow G_c^* = \frac{1}{\tilde{G}} \tag{4.70}$$

This condition is realized when the model is perfect, i.e. if $\tilde{G} = G$.

From the expression

$$Y = G\,U_a + D \tag{4.71}$$

using Eq. (4.69), the general expression of output Y results

$$Y = \frac{G\,G_c^*}{1 + G_c^*(G - \tilde{G})}\,Y_r + \frac{1 - \tilde{G}\,G_c^*}{1 + G_c^*(G - \tilde{G})}\,D \tag{4.72}$$

This expression is equivalent to Eq. (4.63) of direct synthesis if the following controller is chosen

$$G_c = \frac{G_c^*}{1 - \tilde{G}\,G_c^*} \tag{4.73}$$

The disturbance rejection is studied in regulation, so that $Y_r = 0$, and disturbances are perfectly rejected when

$$(1 - \tilde{G}\,G_c^*) = 0 \Longleftrightarrow G_c^* = \frac{1}{\tilde{G}} \tag{4.74}$$

which constitutes a condition analogous to the previous relation (4.70) when the model is perfect.

The theoretical controller is thus given by the transfer function

$$G_c^* = \frac{1}{\tilde{G}} \tag{4.75}$$

This is the classical relation, according to which the controller transfer function is equal to the inverse of the process model. Such a controller would be physically unrealizable when the degree of the denominator of the process transfer

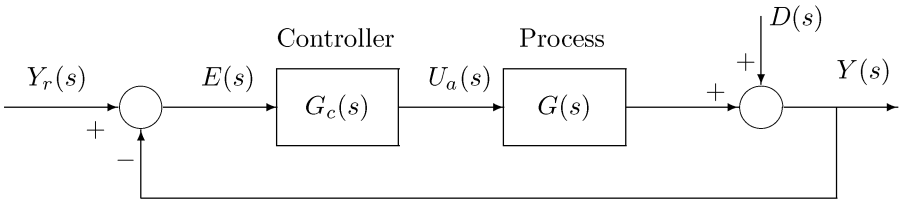


Figure 4.22: Block diagram for classical feedback control

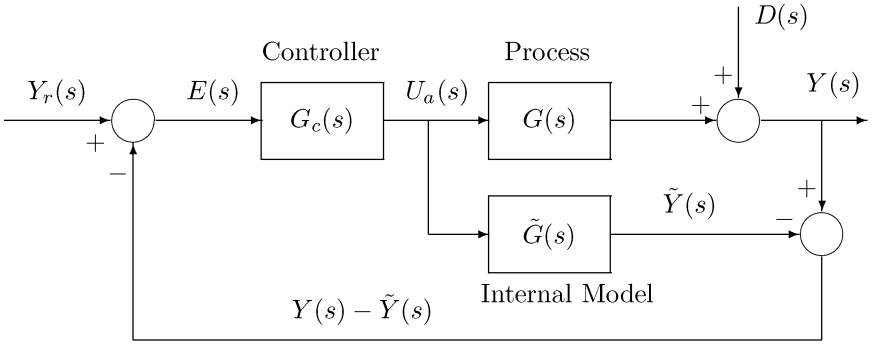


Figure 4.23: Block diagram for internal model control

function is strictly higher than the degree of the numerator of this transfer function. Moreover, positive zeros or zeros with a positive real part as well as time delays present in the process transfer function set a difficulty: positive zeros or zeros with a positive real part for the process would be controller poles and render it unstable. Process time delays would result in pure advances for the controller, thus the latter would be physically unrealizable.

For this reason, internal model control design is performed in two stages: Stage 1: the process model $\tilde{G}$ is decomposed into a product of two factors. The first one, $\tilde{G}_+$ (the gain of which will be taken to be equal to 1) contains time delays and positive zeros or zeros with a positive real part

$$\tilde{G} = \tilde{G}_+ \, \tilde{G}_- \tag{4.76}$$

Stage 2: only $\tilde{G}_-$ is retained (to put aside the time delays and positive zeros or zeros with a positive real part) and the inverse of $\tilde{G}_-$ is filtered (to make the controller physically realizable). The real controller transfer function is then equal to

$$G_c = \frac{1}{\tilde{G}_-} \, f \tag{4.77}$$

where f is a lowpass filter with gain equal to 1.

This filter is typically of the form

$$f = \frac{1}{(\tau_r s + 1)^r} \tag{4.78}$$

where τ_r is the desired closed-loop time constant. The exponent r is chosen so that G_c is a real transfer function (the denominator degree must be larger or equal to that of the numerator) or possibly corresponds to a derivative action (in this case, the denominator degree is lower by one unity to that of the numerator).

With these precautions (decomposition into the two previous factors), the controller G_c is physically realizable and stable. It must be noted that as the internal model control method is based on zeros cancellation, it must not be used for unstable open-loop systems.

In the ideal case where the model is perfect ($\tilde{G} = G$), the output is equal to

$$Y = \tilde{G}_+ f Y_r + (1 - \tilde{G}_+ f)D \tag{4.79}$$

and in a general manner to

$$Y = \frac{Gf}{\tilde{G}_- + f(G - \tilde{G})} Y_r + \frac{\tilde{G}_- - \tilde{G}f}{\tilde{G}_- + f(G - \tilde{G})} D \tag{4.80}$$

For real implementation, it is necessary to consider the individual models of the actuator $\tilde{G}_a$ and the process $\tilde{G}_p$, the measurement $\tilde{G}_m$ such that the previous transfer function $\tilde{G}$ contains the actuator, the process and the measurement

$$\tilde{G} = \tilde{G}_a \tilde{G}_p \tilde{G}_m \tag{4.81}$$

The real implementation will thus be performed according to the block diagram in Fig. 4.24.

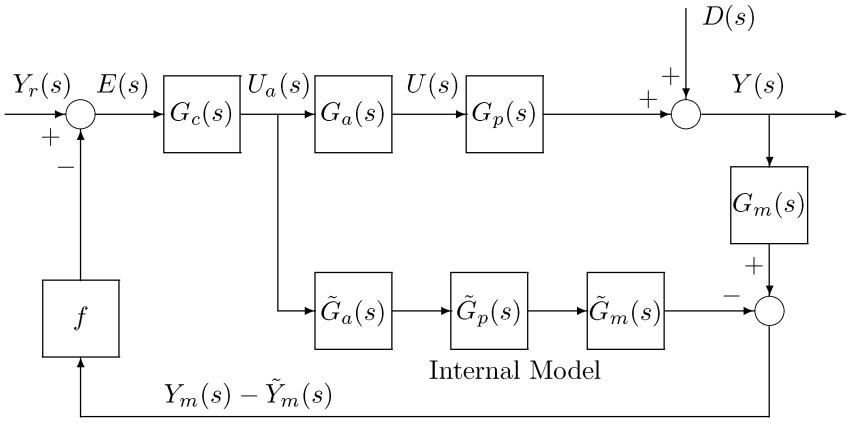


Figure 4.24: Block diagram for real implementation of internal model control

Internal model control allows us to obtain the PID tunings (Rivera et al., 1986) for different process models such as in Table 4.6. In the original table, zero- ($\exp(-t_d s) \approx 1$) or first-order Padé approximations are used for time delays when necessary (the zero-order approximation requires $\epsilon > t_d$ and the

Table 4.6: PID controller parameters based on internal model control from Rivera et al. (1986)

Model	$\frac{Y(s)}{Y_r(s)} = \tilde{G}_+ f$	Controller	$K_c\,K$	τ_I	τ_D
$\frac{K}{\tau s+1}$	$\frac{1}{\epsilon s+1}$	$\frac{\tau s+1}{K\epsilon s}$	$\frac{\tau}{\epsilon}$	τ	
$\frac{K}{(\tau_1 s+1)(\tau_2 s+1)}$	$\frac{1}{\epsilon s+1}$	$\frac{(\tau_1 s+1)(\tau_2 s+1)}{K\epsilon s}$	$\frac{\tau_1+\tau_2}{\epsilon}$	$\tau_1+\tau_2$	$\frac{\tau_1\tau_2}{\tau_1+\tau_2}$
$\frac{K}{\tau^2 s^2+2\zeta\tau s+1}$	$\frac{1}{\epsilon s+1}$	$\frac{\tau^2 s^2+2\zeta\tau s+1}{K\epsilon s}$	$\frac{2\zeta\tau}{\epsilon}$	$2\zeta\tau$	$\frac{\tau}{2\zeta}$
$K\frac{-\beta s+1}{\tau s+1}$	$\frac{-\beta s+1}{\epsilon s+1}$	$\frac{\tau s+1}{K(\beta+\epsilon)s}$	$\frac{\tau}{\beta+\epsilon}$	τ	
$\frac{K}{s}$	$\frac{1}{\epsilon s+1}$	$\frac{1}{K\epsilon}$	$\frac{1}{\epsilon}$	τ	
$K\frac{\exp(-t_d s)}{\tau s+1}$		PI	$\frac{2\tau+t_d}{2\epsilon}$	$\tau+\frac{t_d}{2}$	
$K\frac{\exp(-t_d s)}{\tau s+1}$		PID	$\frac{2\tau+t_d}{2\epsilon+t_d}$	$\tau+\frac{t_d}{2}$	$\frac{\tau t_d}{2\tau+t_d}$

first-order approximation requires $\epsilon > t_d/3$). In the PI and PID tuning with the nonlinear delay term, ϵ should always be greater than 0.1τ and greater than $1.7t_d$ or $0.8t_d$ respectively for PI and PID. ϵ is, in general, the closed-loop time constant that can often be chosen as the dominant open-loop time constant. The original table presents many other models. Comparisons with the Ziegler-Nichols and Cohen-Coon tuning methods in the case of the first-order transfer function with delay are given by Rivera et al. (1986). In general, the performance measure and robustness are better for the IMC design.

Example 4.4: **Internal Model Control of a Thermostatic Bath**

A thermostatic bath (Fig. 4.25) is in charge of maintaining the temperature of a water stream circulating at low velocity by means of a coil in which water vapour circulates. The bath temperature T is the controlled variable while the vapour temperature T_v is the input and the water flow rate F is a modelled disturbance. The process can be approximately represented by the following energy balance

$$F\,C_p\,(T_{in}-T)+U\,A\,(T_v-T)=M\,C_p\,\frac{dT}{dt} \tag{4.82}$$

with the nominal variables and parameters defined in Table 4.7.

By linearizing the system around its steady state, the Laplace model can be written as

$$\bar{T}(s)=\frac{0.0793}{1+2.58\,s}\,\bar{T}_v(s)-\frac{0.331}{1+2.58\,s}\,\bar{F}(s) \tag{4.83}$$

Suppose that, during process identification, errors occur which lead to the

Table 4.7: Nominal variables and parameters of the thermostatic bath

Water flow rate	$F = 25$ kg.min^{-1}
Inlet water temperature	$T_{in} = 294$ K
Water mass to heat	$M = 70$ kg
Global heat transfer coefficient	$UA = 9000$ J.min^{-1}.K^{-1}
Heat capacity of liquid water	$C_p = 4180$ J.kg^{-1}.K^{-1}
Bath temperature	$T = 303$ K
Vapour temperature	$T_v = 407.5$ K

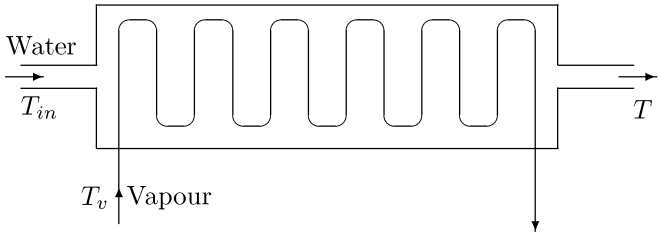


Figure 4.25: Thermostatic bath

following model transfer function

$$\tilde{G}(s) = \frac{0.10}{1 + 3\,s} \tag{4.84}$$

while the disturbance is not measured. This model is used to calculate the controller equal to

$$G_c = \frac{1}{\tilde{G}_-}\, f = \frac{1 + 3\,s}{0.10\,(1 + \tau\,s)} \tag{4.85}$$

In the present case, the filter time constant was chosen to be equal to $\tau = 1$ min. In simulation, instead of a Laplace transfer function for process G, the exact state-space model was used as well as for the influence of the water flow rate disturbance. In this manner, the simulation result is closer to the real process behaviour; it takes into account modelling errors and unmodelled disturbances. After having reached its steady state, the process is first subjected to a set point step at temperature T of 5K at $t = 50$ min, and undergoes flow rate disturbance of 10 kg/min at $t = 100$ min. In Fig. 4.26, it is clear that the set point is reached fast and with very little overshoot, and that the disturbance is rejected without difficulty with a fast comeback to the set point. We notice that the variation of the vapour temperature which is the manipulated input is very steep at the set point step time and that probably in reality it would be limited by saturations or slower dynamics.

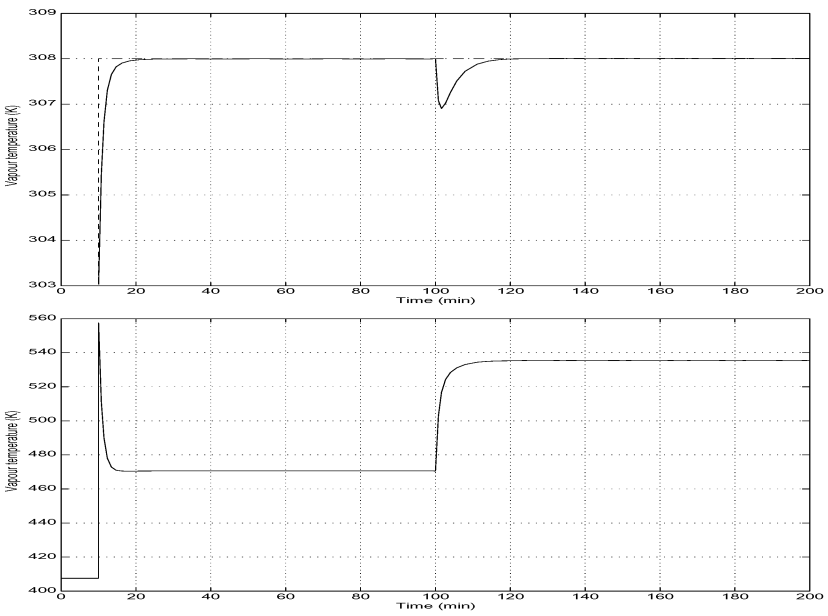


Figure 4.26: Internal model control of a thermostatic bath. Top: set point and bath temperature response to a set point step of 5K at $t = 50\,\text{min}$ and a flow rate disturbance of 10 kg/min at $t = 100\,\text{min}$. Bottom: manipulated input, i.e. vapour temperature

4.9 Pole-Placement

The process transfer function is represented as

$$G(s) = \frac{N(s)}{D(s)} \tag{4.86}$$

with $\deg D \geq \deg N$, and the polynomials $N(s)$ and $D(s)$ are coprime (have no common root); it can be useful to consider that $G(s)$ is strictly proper: the equivalent state-space representation would be controllable and observable; nevertheless this is not compulsory.
The block diagram (Fig. 4.27) can represent any control system such that its linear controller is the most general possible controller. It will be qualified under the general term of pole-placement, which will be justified later, or RST controller. The controller output $u(t)$ is influenced by both controller inputs $y_r(t)$ and $[y(t) + \eta(t)]$, where $\eta(t)$ is the measurement noise. The designation "two-parameter" or "two degrees of freedom" is also used to qualify this controller.

The controller dynamics can be represented by the equation

$$U(s) = \frac{T(s)}{S(s)} Y_r(s) - \frac{R(s)}{S(s)} [Y(s) + \bar{\eta}(s)] \tag{4.87}$$

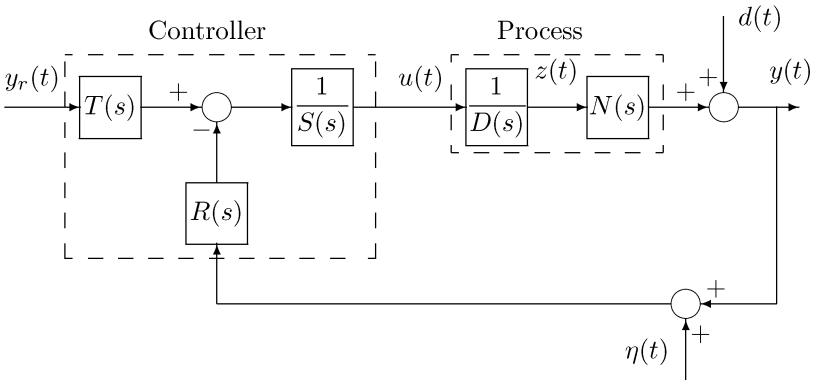


Figure 4.27: Process control system by pole-placement

The controller is supposed to be proper; if it is strictly proper, it will be little influenced by high-frequency noise thus

$$\deg R(s) \leq \deg S(s) \quad \text{and} \qquad \deg T(s) \leq \deg S(s) \tag{4.88}$$

Moreover, $d(t)$ being a disturbance, the output is equal to

$$Y(s) = N(s)\, Z(s) + \bar{d}(s) \tag{4.89}$$

by introducing the partial state $z(t)$. Defining the output error $e(t) = y_r(t) - y(t)$, the important dynamics with respect to exogenous inputs (set point, disturbance, noise) are obtained

$$\begin{bmatrix} Y(s) \\ U(s) \\ E(s) \end{bmatrix} = \frac{1}{DS + NR} \begin{bmatrix} NT & DS & -NR \\ DT & -DR & -DR \\ DS + NR - NT & -DS & NR \end{bmatrix} \begin{bmatrix} Y_r(s) \\ \bar{d}(s) \\ \bar{\eta}(s) \end{bmatrix} \tag{4.90}$$

Closed-loop stability is thus conditioned by the roots of polynomial $P(s)$ of Bezout equation or Diophantine equation[1]

$$P(s) = D(s)\, S(s) + N(s)\, R(s) \tag{4.91}$$

which are the closed-loop transfer function poles, thus impose the regulation and tracking behaviour. Only the polynomials N and D are known

$$\begin{aligned} D(s) &= D_0 s^n + D_1 s^{n-1} + \ldots + D_{n-1} s + D_n \\ N(s) &= N_0 s^n + N_1 s^{n-1} + \ldots + N_{n-1} s + N_n \end{aligned} \tag{4.92}$$

with D and N coprime. Because of hypotheses concerning proper transfer functions, we obtain

$$\deg P(s) = \deg D\, S = n + \deg S \tag{4.93}$$

[1]From the name of Greek mathematician Diophante (325-410), also called the Bezout equation.

where n is the process order.

Set

$$\begin{aligned} S(s) &= S_0 s^{n_s} + S_1 s^{n_s-1} + \ldots + S_{n_s-1} s + S_{n_s} \\ R(s) &= R_0 s^{n_r} + R_1 s^{n_r-1} + \ldots + R_{n_r-1} s + R_{n_r} \end{aligned} \tag{4.94}$$

The model of the closed-loop dynamics is imposed by the user by specifying the closed-loop transfer function, thus the poles (hence the name pole-placement)

$$\frac{Y(s)}{Y_r(s)} = G_m(s) = \frac{N_m(s)}{D_m(s)} \tag{4.95}$$

where $N_m(s)$ and $D_m(s)$ are coprime and $D_m(s)$ fixes the closed-loop poles. To get the closed-loop poles, $D_m(s)$ must divide $P(s)$.

It is possible to distinguish:

- The placement of poles and zeros where, at the same time, poles and zeros of the closed-loop transfer function are specified; it is also called reference model control.
- The pole-placement where only the poles of the closed-loop transfer function are placed.

These two cases will be considered simultaneously, and the consequences of this distinction will be studied.

The process zeros, corresponding to $N(s) = 0$, are kept in closed loop, except if they are specifically cancelled by corresponding poles. Aström and Wittenmark (1989) recommend that the choice of $G_m(s)$ be related to that of G; although the zeros of $G(s)$ can be modified by pole-zero cancellation, Aström and Wittenmark (1989) advise keeping the zeros in the model $G_m(s)$. Unstable and weakly damped zeros cannot be cancelled and $N(s)$ is factorized as

$$N(s) = N^+(s)\, N^-(s) \tag{4.96}$$

where $N^+(s)$ is a monic[2] polynomial containing the factors that can be eliminated (corresponding to stable and well-damped zeros) and $N^-(s)$ contains the remaining factors. Thus, the polynomial $P(s)$ must be divisible by $N^+(s)$. De Larminat (1993) proposes a strategy that realizes a robust pole-placement (Fig. 4.28) setting the zeros of $P(s)$ from the zeros of $N(s)$. The recommended procedure is developed according to the following stages (if necessary):

a) Any unstable zero of $N(s)$ is first replaced by its symmetrical zero with respect to the imaginary axis ($z_1 \rightarrow z_1'$).

b) A weakly damped complex zero is replaced by a conveniently damped zero ($z_2 \rightarrow z_2'$).

c) A slow zero is replaced by a zero of convenient dynamics (removal from the imaginary axis to a distance reciprocal of the settling time) ($z_3 \rightarrow z_3'$).

[2] A polynomial $P(x)$ is called monic if the coefficient of highest degree monomial is equal to 1

$$P(x) = x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n$$

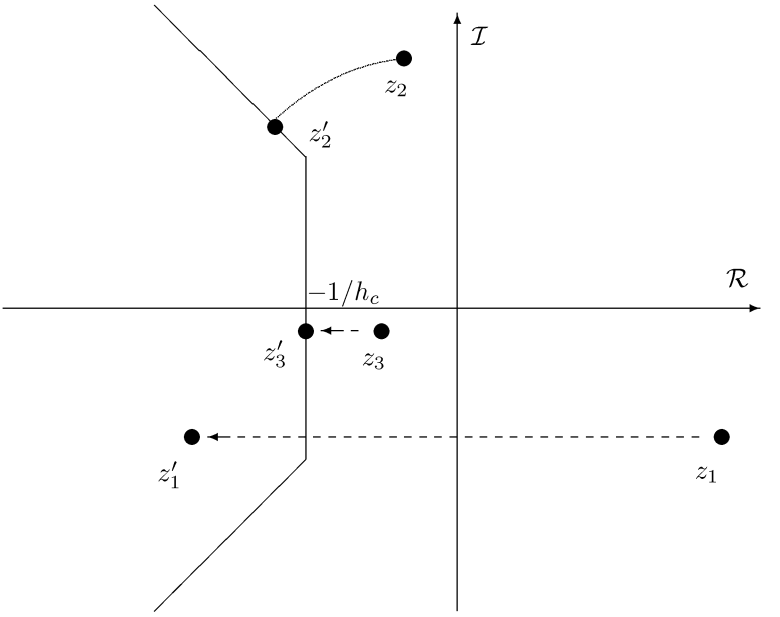


Figure 4.28: Determination of the zeros of $P(s)$ from the zeros of $N(s)$

In general, the polynomial $P(s)$ has a degree higher than $D_m(s)N^+(s)$, which divides it, and one sets

$$P(s) = D_m(s)\, N^+(s)\, D_o(s) \tag{4.97}$$

where $D_o(s)$ qualifies the observer dynamics for polynomial $P(s)$: the factor $D_o(s)$ cancels itself by pole-zero simplification and this dynamics is not controllable by y_r. $D_o(s)$ will have to be stable and will have a faster dynamics than that of $D_m(s)$.

The polynomial $P(s)$ is now specified by Eq. (4.97). As, moreover, $N(s) = N^+(s)N^-(s)$, the Bezout equation

$$D(s)\, S(s) + N^+(s)\, N^-(s)\, R(s) = D_m(s)\, N^+(s)\, D_o(s) \tag{4.98}$$

imposes that $N^+(s)$ divides $S(s)$ (because $N^+(s)$ is coprime with $D(s)$), thus

$$S(s) = N^+(s)\, S_1(s) \tag{4.99}$$

so that the Bezout equation takes the definitive form

$$D(s)\, S_1(s) + N^-(s)\, R(s) = D_m(s)\, D_o(s) \tag{4.100}$$

From the expressions of the closed-loop transfer function, one deduces

$$T(s)=\frac{P(s)}{N(s)}\,\frac{Y(s)}{Y_r(s)} = \frac{P(s)}{N(s)}\,\frac{N_m(s)}{D_m(s)}=\frac{D_m(s)\, N^+(s)\, D_o(s)}{N^+(s)\, N^-(s)}\,\frac{N_m(s)}{D_m(s)}=\frac{D_o(s)}{N^-(s)}\, N_m(s) \tag{4.101}$$

In order that $T(s)$ is a polynomial and the Bezout equation has a solution, it is necessary that $N^-(s)$ divides $N_m(s)$, thus

$$N_m(s) = N^-(s)\,N_{m,1}(s) \tag{4.102}$$

One obtains

$$T(s) = D_o(s)\,N_{m,1}(s) \tag{4.103}$$

The conditions on the polynomial degrees (with $\deg D = n$) are

$$\begin{array}{l} \deg D_o \geq 2n - \deg D_m - \deg N^+ - 1 \\ \deg D_m - \deg N_m \geq \deg D - \deg N \end{array} \tag{4.104}$$

thus allowing the physical realizability of the controller.

The control law is deduced from polynomials (R, S, T)

$$S(s)\,U(s) = T(s)\,Y_r(s) - R(s)\,Y(s) \tag{4.105}$$

Pole-placement, in the most general formulation, that is to say with partial zeros simplification (stable and well-damped), is thus constituted by Eqs. (4.95), (4.96), (4.97), (4.99), (4.100), (4.102), (4.103), (4.104) and (4.105).

In the case where all zeros are simplified, it suffices to take $N^+(s) = N(s)/\alpha$ and $N^-(s) = \alpha$, α being scalar, with the same equations as previously.

In the case where there is no zero simplification, it suffices to take $N^- = N(s)$ and $N^+(s) = 1$.

The solution of a general Bezout equation of the form

$$A\,S + B\,R = P \tag{4.106}$$

$$\begin{array}{ll} A(s) & = a_0 s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n \\ B(s) & = b_0 s^m + b_1 s^{m-1} + \ldots + b_{m-1} s + b_m \\ P(s) & = P_0 s^p + P_1 s^{p-1} + \ldots + P_{p-1} s + P_p \\ R(s) & = R_0 s^r + R_1 s^{r-1} + \ldots + R_{r-1} s + R_r \\ S(s) & = S_0 s^k + S_1 s^{k-1} + \ldots + S_{k-1} s + S_k \end{array} \tag{4.107}$$

is performed by identification on successive powers of $P(s)$ $(p = k + r + 2)$; a linear set of equations results

$$\left[\begin{array}{ccccc|cccc}
a_0 & 0 & \ldots & & 0 & 0 & \ldots & & 0 \\
a_1 & a_0 & & & \vdots & \vdots & & & \vdots \\
\vdots & & \ddots & & 0 & b_0 & & & \\
 & & & & & b_1 & b_0 & & \\
a_n & \ldots & & & a_0 & \vdots & & \ddots & 0 \\
0 & \ddots & & & \vdots & b_n & & & b_0 \\
\vdots & & \ddots & & \vdots & 0 & \ddots & & \vdots \\
 & & & a_n & a_{n-1} & \vdots & & b_n & b_{n-1} \\
0 & \ldots & & 0 & a_n & 0 & \ldots & 0 & b_n
\end{array}\right]
\begin{bmatrix} S_0 \\ \vdots \\ S_k \\ R_0 \\ \vdots \\ R_r \end{bmatrix}
=
\begin{bmatrix} P_0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ P_p \end{bmatrix}$$

$$\underbrace{\qquad\qquad\qquad}_{k+1 \text{ columns}} \quad \underbrace{\qquad\qquad\qquad}_{r+1 \text{ columns}} \tag{4.108}$$

The matrix to be inverted during the solution is a Sylvester matrix, of dimension $(2n \times 2n)$, if $k = r = n - 1$. For a Sylvester matrix to be nonsingular, thus invertible, a necessary and sufficient condition is that polynomials $A(s)$ and $B(s)$ are coprime. Note that if $S(s)$ was supposed to be monic, this would remove an unknown in the problem and would impose a constraint on $P(s)$.

In this design of a pole-placement controller, there always exists a pole-zero simplification. The pole-placement controller must be designed as a set that receives the reference $y_r(t)$ and the output $y(t)$ and returns the input $u(t)$, according to the diagram in Fig. 4.27. Chen (1993) proposes an analogous implementation from the equivalent state-space scheme. Rather than the theoretical scheme of the pole-placement controller (Fig. 4.27), De Larminat (1993) recommends the scheme in Fig. 4.29. It is easy to show that this scheme is equivalent by setting

$$T(s) = \frac{R(0)}{C(0)}\, C(s) \tag{4.109}$$

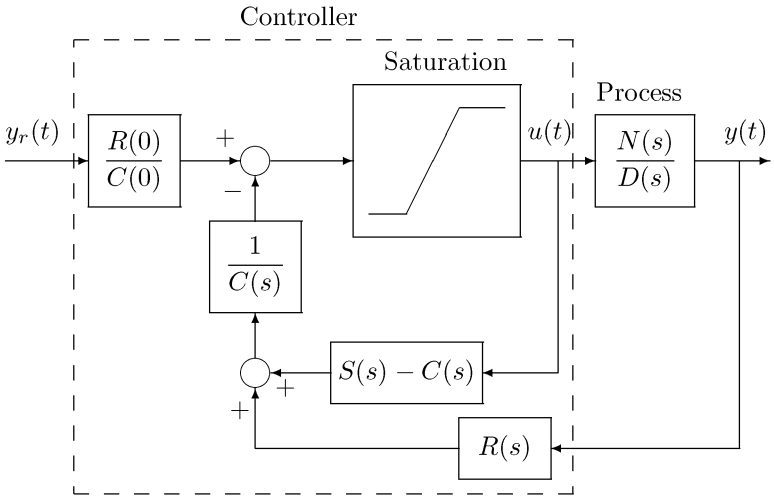


Figure 4.29: Recommended scheme (De Larminat, 1993) for the implementation of a pole-placement controller

Wolovich (1994) shows that although the choice of the observer polynomial $D_o(s)$ has no apparent influence on the closed-loop response, as the closed-loop transfer function is identical whatever $D_o(s)$ is, indeed the choice of $D_o(s)$ influences robustness by modifying the sensitivity function, which takes into account process model uncertainties.

Example 4.5: **Pole-Placement for a Continuous Transfer Function**

Consider the following open-loop unstable process having the transfer function

$$G(s) = \frac{5}{(s-1)(s+2)} \tag{4.110}$$

giving: $N(s) = 5$ and $D(s) = s^2 + s - 2$. According to the ITAE criterion (Table 4.1), the desired closed-loop transfer function is chosen, which gives the step response of Fig. 4.30

$$G_m(s) = \frac{\omega_0^2}{s^2 + 1.4\omega_0 s + \omega_0^2} \tag{4.111}$$

with $\omega_0 = 0.5$, corresponding to $N_m = 0.25$ and $D_m = s^2 + 0.7s + 0.25$. We deduce $N^+(s) = 1$, $N^-(s) = 5$, $N_{m,1} = 0.05$ and $S_1 = S$.

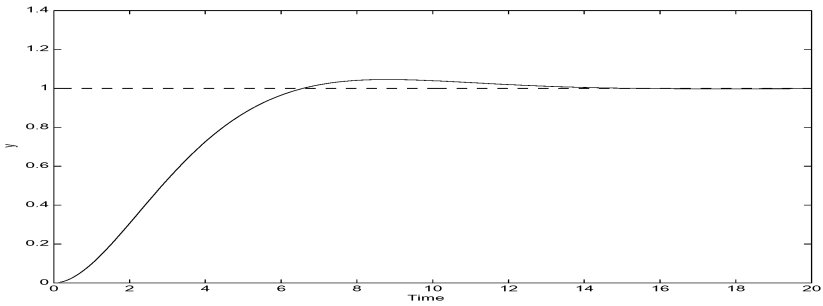


Figure 4.30: Pole-placement controller: response to a set point unit step

The polynomial $D_o(s)$ which will cancel itself by pole-zero simplification is chosen to be equal to $D_o(s) = s + 5$, hence

$$\begin{aligned} P(s) &= D_m(s)\, N^+(s)\, D_o(s) \\ &= s^3 + 5.7s^2 + 3.75s + 1.25 \end{aligned} \tag{4.112}$$

with, furthermore,

$$\begin{aligned} P(s) &= N^-(s)R(s) + D(s)S_1(s) \\ &= 5R(s) + (s^2 + s - 2)S(s) \end{aligned} \tag{4.113}$$

The degrees of R and S are deduced: $\deg R = 1$ and $\deg S = 1$. The Sylvester matrix is then of dimension (4×4) and the system to be solved corresponding to the Bezout equation is written as

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -2 & 1 & 5 & 0 \\ 0 & -2 & 0 & 5 \end{bmatrix} \begin{bmatrix} S_0 \\ S_1 \\ R_0 \\ R_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 5.7 \\ 3.75 \\ 1.25 \end{bmatrix} \tag{4.114}$$

The polynomials R and S are the solutions of the previous Bezout equation

$$R(s) = 0.21s + 2.13\ ; \qquad S(s) = S_1(s) = s + 4.7 \tag{4.115}$$

From Eq. (4.103), the polynomial T is obtained

$$T(s) = D_o(s)N_{m,1}(s) = (s + 5)0.05 = 0.05s + 0.25 \tag{4.116}$$

The closed-loop behaviour will be that of the specified transfer function $G_m(s)$ (Fig. 4.30). Note that the overshoot is very low.

4.9.1 Robustness of Pole-Placement Control

Equation (4.90) gives the output error with respect to set point or reference, disturbance and noise

$$E(s) = \frac{1}{DS + NR} \left\{ (DS + NR - NT)\, Y_r(s) - DS\, \bar{d}(s) + NR\, \bar{\eta}(s) \right\} \quad (4.117)$$

The set point can be expressed as

$$Y_r(s) = \frac{N^c(s)}{D^c(s)} \quad (4.118)$$

The output error with respect to the set point can be decomposed into the natural error $e^n(t)$ and the forced error $e^f(t)$, related to the type of set point

$$\begin{aligned} E(s) &= \frac{DS + NR - NT}{DS + NR} \frac{N^c(s)}{D^c(s)} = \frac{N^e(s)}{P(s)} \frac{N^c(s)}{D^c(s)} \\ &= \frac{N^{en}}{DS + NR} + \frac{N^{ef}(s)}{D^c(s)} \\ &= E^n(s) + E^f(s) \end{aligned} \quad (4.119)$$

A means to control the steady-state error is to fulfil the conditions so that the forced error be zero, thus $N^{ef}(s) = 0$ as the natural error tends towards 0 when $t \to \infty$.

The problem is now to find how we can obtain $N^{ef}(s) = 0$. It suffices that

$$\frac{N^e\, N^c}{P(s)\, D^c(s)} = \frac{N^{en}}{P} \quad (4.120)$$

This implies that $D^c(s)$ must divide $N^e(s)$ as $D^c(s)$ is coprime with $N^c(s)$. Set

$$N^e(s) = N^{e'}(s)\, D^c(s) \quad (4.121)$$

which would give the error

$$E(s) = \frac{N^c(s)\, N^{e'}(s)}{P(s)} \quad (4.122)$$

In pole-placement, the internal model principle implies that a model of $D^c(s)$, the denominator in the expression (4.119) of the forced error, is present in the denominator $D(s)S(s)$ of the open-loop transfer function, thus

$$D(s)S(s) = N^{e'}(s)\, D^c(s) \quad (4.123)$$

on the other hand, if

$$R(s) - T(s) = R'(s)\, D^c(s) \tag{4.124}$$

these two equations imply that the error

$$\begin{aligned} E(s) &= \frac{DS + NR - NT}{DS + NR} \frac{N^c(s)}{D^c(s)} \\ &= \frac{N^{e'} D^c + R' D^c}{DS + NR} \frac{N^c(s)}{D^c(s)} \\ &= \frac{N^c(N^{e'} + R')}{DS + NR} \end{aligned} \tag{4.125}$$

does not depend on D^c, thus the forced error is zero. Conditions (4.123) and (4.124) thus ensure robustness with respect to eventual process variations modifying polynomials $N(s)$ and $D(s)$ (Wolovich, 1994).

Similarly, when the dynamics of undamped disturbances $d(t)$ is known, e.g. $\bar{d}(s) = 1/s$ for a step, it is advised to take it into account in the controller by including the part of the denominator of $\bar{d}(s)$ that poses a problem in the product $D(s)S(s)$ (refer to Eq. (4.90)), in accordance with the internal model principle. For the disturbance, this would mean that s should be a divider of $D(s)S(s)$, thus, in general, of $S(s)$.

4.9.2 Unitary Feedback Controller

The general pole-placement controller of Fig. 4.27 can be reduced under the form of a unitary feedback controller, also called a one degree of freedom controller (Fig. 4.31) by assuming that $T(s) = R(s)$. The controller transfer function becomes simply equal to

$$G_c(s) = \frac{R(s)}{S(s)} \tag{4.126}$$

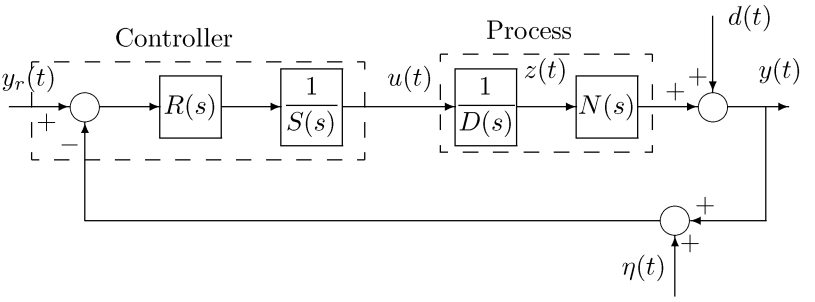


Figure 4.31: Process control system with a one degree of freedom controller

For this controller, the error $e(t) = y_r(t) - y(t)$ rules the controller and is made zero by the output unitary feedback.

The fact that the controller has one or two degrees of freedom does not modify the number of integrators as, in both cases, the controller order is equal

to the order of $S(s)$. Of course, it is simpler to design the one degree of freedom controller, as it is sufficient to determine the transfer function $R(s)/S(s)$. The one degree of freedom controller makes only a pole-placement.

4.10 Linear Quadratic Control

In this section, a voluntarily reduced presentation of linear quadratic control is done by the use of continuous transfer functions (Wolovich, 1994). A more complete presentation, in multivariable state space, in continuous-time and in discrete-time, is outlined in Chap. 14 in relation to optimal control.

Linear quadratic control allows us to choose a quadratic criterion which takes into account not only the error, but also the control signal energy

$$J_{lq} = \int_0^\infty \{[y(t) - y_r(t)]^2 + \rho u^2(t)\} dt \tag{4.127}$$

where ρ is a positive constant. The problem is then to find a global transfer function which minimizes this criterion with respect to the control $u(t)$. Assume that the process transfer function is of the form

$$G(s) = \frac{N(s)}{D(s)} \tag{4.128}$$

with $\deg D \geq \deg N$, and polynomials $N(s)$ and $D(s)$ being coprime (having no common roots). The polynomial

$$\Delta(s) = \rho\, D(s)\, D(-s) + N(s)\, N(-s) \tag{4.129}$$

can only take strictly positive values (for $\rho \geq 0$) when s varies in the s-plane. The roots of $\Delta(s)$ are symmetrical with respect to the imaginary axis. The spectral factorization can be performed such as

$$\Delta(s) = \Delta^+(s)\, \Delta^-(s) \tag{4.130}$$

where $\Delta^+(s)$ contains all roots of $\Delta(s)$ located in the complex left half-plane and $\Delta^-(s)$ in the complex right half-plane. Spectral factorization can be considered in the frequency domain as the dual of the Kalman filter in the time domain (Middleton and Goodwin, 1990)

Denoting the derivative operator associated with the Laplace variable s by δ, from the process transfer function, neglecting the disturbances, the input can be written as

$$u(t) = D(\delta)\, z(t)\ ; \quad y(t) = N(\delta)\, z(t) \tag{4.131}$$

where $z(t)$ is the partial state (as in Fig. 4.31).

4.10.1 Regulation Behaviour

The criterion (4.127) then becomes

$$J_{lq} = \int_0^\infty [y^2(t) + \rho u^2(t)]dt = \int_0^\infty [N^2(\delta) + \rho D^2(\delta)]z^2(t)dt \tag{4.132}$$

An important property is that the integral

$$\int_0^\infty [N^2(\delta) + \rho D^2(\delta) - \Delta^{+^2}(\delta)]z^2(t)dt \tag{4.133}$$

does not depend on the path followed by $z(t)$. It results that the criterion presented in the form

$$J_{lq} = \int_0^\infty [N^2(\delta) + \rho D^2(\delta) - \Delta^{+^2}(\delta)]z^2(t)dt + \int_0^\infty \Delta^{+^2}(\delta)\, z^2(t)dt \tag{4.134}$$

is minimum if

$$\Delta^{+^2}(\delta)\, z^2(t) = 0 \Longleftrightarrow \Delta^{+}(\delta)\, z(t) = 0 \tag{4.135}$$

hence the optimal control law

$$u^*(t) = [D(\delta) - \Delta^{+}(\delta)]\, z(t) \tag{4.136}$$

The closed-loop transfer function is

$$G_{cl}(s) = \frac{Y(s)}{Y_r(s)} = \frac{N(s)\, T(s)}{D(s)\, S(s) + N(s)\, R(s)} = \frac{N(s)\, \alpha\, T'(s)}{\Delta^{+}(s) T'(s)} = \frac{N(s)\, \alpha}{\Delta^{+}(s)} \tag{4.137}$$

so that the closed-loop poles are the roots of $\Delta^{+}(s)$. From Eq. (4.137), α is a scalar such that $T(s) = \alpha T'(s)$.

4.10.2 Tracking Behaviour

To ensure zero steady-state error, the robustness equations of pole-placement are used, considering that quadratic control constitutes an optimal pole-placement. In particular, the set point is given by

$$Y_r(s) = \frac{N^c(s)}{D^c(s)} \tag{4.138}$$

and the following equations of pole-placement are valid

$$D(s)S(s) = N^{e'}(s)\, D^c(s)\ ; \qquad R(s) - T(s) = R'(s)\, D^c(s) \tag{4.139}$$

The monic polynomial $D^m(s)$ is defined as the smallest common multiple of $D(s)$ and $D^c(s)$ so that we have simultaneously

$$\begin{array}{ll} D^m(s) = D(s)\, D^{c'}(s)\ ; & D^m(s) = D'(s)\, D^c(s) \\ \deg D^{c'}(s) \leq \deg D^c(s)\ ; & \deg D'(s) \leq \deg D(s) \end{array} \tag{4.140}$$

Supposing that $D^m(s)$ and $N(s)$ are coprime, an extended spectral factorization is realized

$$\begin{aligned}\Delta^e(s) &= \rho\, D^m(s)\, D^m(-s) + N(s)\, N(-s)\\ &= \Delta^{e+}(s)\, \Delta^{e-}(s)\end{aligned} \tag{4.141}$$

where $\Delta^{e+}(s)$ contains all roots of $\Delta^e(s)$ located in the complex left half-plane and $\Delta^{e-}(s)$ in the right half-plane.

By duality, an optimal observer can be defined from the spectral factorization

$$\begin{aligned}\Delta^o(s) &= \sigma\, D(s)\, D(-s) + N(s)\, N(-s)\\ &= \Delta^{o+}(s)\, \Delta^{o-}(s)\end{aligned} \tag{4.142}$$

where σ qualifies noise intensity. Let $T_1(s)$ be a monic polynomial formed with any $n-1$ roots of $\Delta^{o+}(s)$ (nevertheless, the $n-1$ fastest or slowest roots can be chosen). By choosing the slowest roots (closer to the imaginary axis) which will cancel by pole-zero simplification, the closed-loop response is faster. It is also possible to proceed differently, as is shown in the second part of the example.

Pole-placement will be specified by the polynomials $R(s)$, $S(s) = D^{c'}(s)\, S'(s)$ and $T(s) = T_1(s)T'(s)$.

A first Bezout equation is then obtained, which gives $R(s)$ and $S'(s)$:

$$\begin{aligned}&D(s)\, S(s) + N(s)\, R(s) = P(s) \Longrightarrow\\ &D(s)\, D^{c'}(s)\, S'(s) + N(s)\, R(s) = \Delta^{e+}(s)\, T_1(s)\end{aligned} \tag{4.143}$$

then a second Bezout equation is deduced from Eq. (4.139), which gives $T'(s)$ and $R'(s)$:

$$T_1(s)\, T'(s) + D^c(s)\, R'(s) = R(s) \tag{4.144}$$

The robustness condition $D(s)S(s) = N^{e'}(s)\, D^c(s)$ is fulfilled, as

$$D(s)S(s) = D(s)\, D^{c'}(s)\, S'(s) \tag{4.145}$$

contains all the roots of $D^c(s)$.

On the other hand, the closed-loop transfer function is equal to

$$G_{cl}(s) = \frac{Y(s)}{Y_r(s)} = \frac{N(s)\, T(s)}{D(s)S(s) + N(s)R(s)} = \frac{N(s)\, T_1(s)\, T'(s)}{\Delta^{e+}\, T_1(s)} = \frac{N(s)\, T'(s)}{\Delta^{e+}} \tag{4.146}$$

an expression analogous to Eq. (4.137).

Example 4.6: **Linear Quadratic Control for a Continuous Transfer Function**

The same open-loop unstable system as for pole-placement is considered with transfer function $G(s)$ being equal to

$$G(s) = \frac{5}{(s-1)(s+2)} = \frac{N(s)}{D(s)} \tag{4.147}$$

with $N(s) = 5$, $D(s) = (s-1)(s+2)$.

The user's wish is that the control system allows us to realize a tracking when the set point is step-like. This corresponds to

$$Y_r(s) = \frac{1}{s} = \frac{N^c(s)}{D^c(s)} \tag{4.148}$$

with $N^c(s) = 1$, $D^c(s) = s$. The monic polynomial $D^m(s)$, the smallest common multiple of $D(s)$ and $D^c(s)$ is thus equal to

$$D^m(s) = s(s-1)(s+2) \tag{4.149}$$

with $D^{c'}(s) = s$.

The extended spectral factorization gives

$$\begin{aligned} \Delta^e(s) &= \rho\, D^m(s)\, D^m(-s) + N(s)\, N(-s) \\ &= -\rho s^6 + 5\rho s^4 - 4\rho s^2 + 25 \\ &= \Delta^{e+}(s)\, \Delta^{e-}(s) \end{aligned} \tag{4.150}$$

hence, by choosing $\rho = 1$,

$$\Delta^{e+}(s) = s^3 + 4.3308 s^2 + 6.878 s + 5 \tag{4.151}$$

has the following roots: $(-1.0293 \pm 1.0682i)$ and -2.2722.

The spectral factorization for the observer gives

$$\begin{aligned} \Delta^o(s) &= \sigma\, D(s)\, D(-s) + N(s)\, N(-s) \\ &= \sigma s^4 - 5\sigma s^2 + 4\sigma + 25 \\ &= \Delta^{o+}(s)\, \Delta^{o-}(s) \end{aligned} \tag{4.152}$$

hence, by choosing $\sigma = 1$

$$\Delta^{o+}(s) = s^2 + 3.9712 s + 5.3852 \tag{4.153}$$

has the following roots: $(-1.9856 \pm 1.2011i)$.

Wolovich (1994) recommends to choose for $T_1(s)$ a monic polynomial formed with any $n-1$ roots of $\Delta^{o+}(s)$. Here, $n = 2$ and the conjugate complex roots do not allow this choice. We have simply chosen $T_1 = \Delta^{o+}(s)$.

It is then possible to solve the first Bezout equation

$$\begin{aligned} &D(s)\, D^{c'}(s)\, S'(s) + N(s)\, R(s) = \Delta^{e+}(s)\, T_1(s) \\ &s(s-1)(s+2)\, S'(s) + 5\, R(s) = \\ &(s^3 + 4.3308 s^2 + 6.878 s + 5)\, (s^2 + 3.9712 s + 5.3852) \end{aligned} \tag{4.154}$$

which gives

$$R(s) = 9.2162 s^2 + 21.0431 s + 5.3852\,; \qquad S'(s) = s^2 + 7.302 s + 24.1598 \tag{4.155}$$

Two different cases have been studied in the following :
First case:

The second Bezout equation is solved

$$\begin{aligned} &T_1(s)\,T'(s) + D^c(s)\,R'(s) = R(s) \\ &(s^2 + 3.9712s + 5.3852)\,T'(s) + s\,R'(s) = 9.2162s^2 + 21.0431s + 5.3852 \end{aligned} \tag{4.156}$$

which gives

$$T'(s) = 1\ ; \qquad R'(s) = 8.2162s + 17.0719 \tag{4.157}$$

The polynomials of the equivalent pole-placement are thus

$$\begin{aligned} R(s) &= 9.2162s^2 + 21.0431s + 5.3852 \\ S(s) &= D^{c'}(s)\,S'(s) = s^3 + 7.302s^2 + 24.1598s \\ T(s) &= T_1\,T'(s) = s^2 + 3.9712s + 5.3852 \end{aligned} \tag{4.158}$$

The resulting closed-loop transfer function is equal to

$$\begin{aligned} G_{cl}(s) &= \frac{Y(s)}{Y_r(s)} = \frac{NT}{DS + NR} = \frac{NT_1T'}{\Delta^{e+}T_1} = \frac{NT'}{\Delta^{e+}} \\ &= \frac{5}{s^3 + 4.3308s^2 + 6.878s + 5} \end{aligned} \tag{4.159}$$

the very good step response of which is shown in Fig. 4.32 and presents a very slight overshoot.

Second case:

The polynomial $T_1(s)$, which will be used in the second Bezout equation, is replaced by $\tilde{T}_1(s)$ obtained from the $n-1$ slowest roots of $\Delta^{e+}(s)\Delta^{o+}(s)$. In fact, in the present case, two complex roots are chosen, resulting in

$$\tilde{T}_1(s) = (s + 1.0293 + 1.0682i)(s + 1.0293 - 1.0682i) = s^2 + 2.0586s + 2.2005 \tag{4.160}$$

The second Bezout equation is solved

$$\begin{aligned} &\tilde{T}_1(s)\,T'(s) + D^c(s)\,R'(s) = R(s) \\ &(s^2 + 2.0586s + 2.2005)\,T'(s) + s\,R'(s) = 9.2162s^2 + 21.0431s + 5.3852 \end{aligned} \tag{4.161}$$

which gives

$$T'(s) = 2.4472\ ; \qquad R'(s) = 6.7689s + 16.0052 \tag{4.162}$$

The polynomials of the equivalent pole-placement are thus

$$\begin{aligned} R(s) &= 9.2162s^2 + 21.0431s + 5.3852 \\ S(s) &= D^{c'}(s)\,S'(s) = s^3 + 7.302s^2 + 24.1598s \\ T(s) &= \tilde{T}_1(s)\,T'(s) = 2.4472s^2 + 5.0379s + 5.3852 \end{aligned} \tag{4.163}$$

The resulting closed-loop transfer function is equal to

$$\begin{aligned} G_{cl}(s) &= \frac{Y(s)}{Y_r(s)} = \frac{NT}{DS + NR} = \frac{N\tilde{T}_1\,T'}{\Delta^{e+}T_1} \\ &= \frac{12.2360}{s^3 + 6.2434s^2 + 14.409s + 12.2360} \end{aligned} \tag{4.164}$$

the step response of which (faster and without any overshoot compared to the first case) is shown in Fig. 4.32. It is also possible to compare the obtained curves with the curve obtained from the pole-placement with reference model (Fig. 4.30).

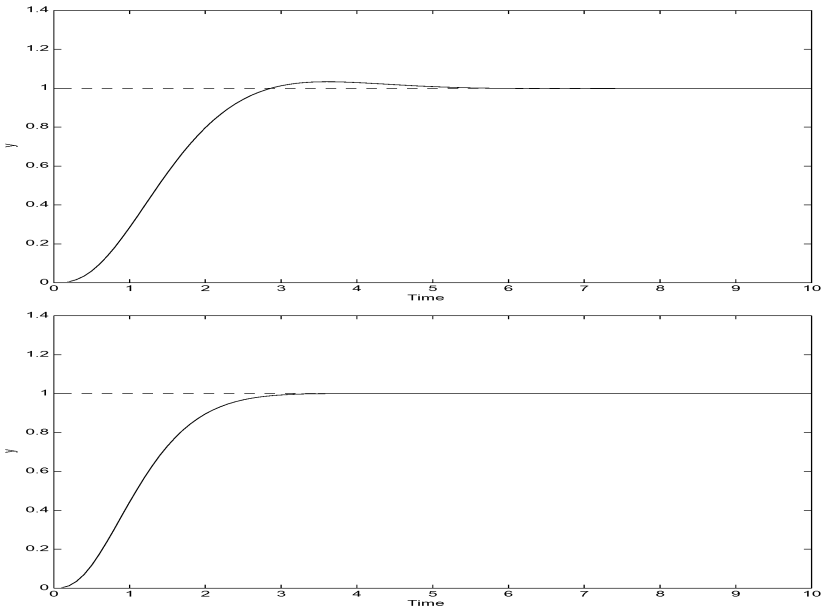


Figure 4.32: Polynomial linear quadratic control: response to a set point unit step in the two cases (first case at top, second case at bottom)

References

K.J. Aström and T. Hagglund. Revisiting the ziegler-nichols step response method for pid control. *J. Proc. Cont.*, 14:635–650, 2004.

K.J. Aström and T. Hägglund. *Automatic Tuning of PID Controllers*. Instrument Society of America, 1988.

K.J. Aström and B. Wittenmark. *Adaptive Control*. Addison-Wesley, New York, 1989.

C.T. Chen. *Analog and Digital Control System Design: Transfer-Function, State-Space, and Algebraic Methods*. Harcourt Brace Jovanovich College, Fort Worth, 1993.

J.Y. Choi, H.G. Pandit, R.R. Rhinehart, and R.J. Farrell. A process simulator for pH control studies. *Comp. Chem. Engng.*, 19:527–539, 1995.

G.H. Cohen and G.A. Coon. Theoretical considerations of retarded control. *Trans. ASME*, 75:827–834, 1953.

D.R. Coughanowr and L.B. Koppel. *Process Systems Analysis and Control*. McGraw-Hill, Auckland, 1985.

P. De Larminat. *Automatique, Commande des Systèmes Linéaires*. Hermès, Paris, 1993.

B.A. Francis and W.M. Wonham. The internal model principle of control theory. *Automatica*, 12:457–465, 1976.

C.E. Garcia and M. Morari. Internal model control. 1. A unifying review and some new results. *Ind. Eng. Chem. Process Des. Dev.*, 21:308–323, 1982.

D. Graham and R.C. Lathrop. The synthesis of optimum response: criteria and standard forms. *AIEE Transactions Part II*, 72:273–288, 1953.

S.R. Gupta and D.R. Coughanowr. On-line gain identification of flow processes with application to adaptive pH control. *AIChE J.*, 24:654–664, 1978.

T.K. Gustafsson and K.V. Waller. Dynamic modeling and reaction invariant control of pH. *Chem. Eng. Sci.*, 38:389, 1983.

T.K. Gustafsson and K.V. Waller. Nonlinear and adaptive control of pH. *Ind. Eng. Chem. Research*, 31:2681, 1992.

T.K. Gustafsson, B.O. Skrifvars, K.V. Sandstrom, and K.V. Waller. Modeling of pH for control. *Ind. Eng. Chem. Research*, 34:820, 1995.

C.C. Hang, K.J. Aström, and W.K. Ho. Refinements of the ziegler-nichols tuning formulas. *IEE Proceeding-D*, 138(2):111–118, 1991.

R. Hanus. Le conditionnement. Une méthode générale d'antichoc des régulateurs. *APII*, 24:171–186, 1990.

R. Hanus. Systèmes d'anti-emballement des régulateurs. In A. Rachid, editor, *Systèmes de Régulation*, pages 54–83. Masson, Paris, 1996.

R. Hanus and Y. Peng. Conditioning technique for controllers with time delay. *IEEE Trans. Automat. Control*, AC-37:689–692, 1992.

R. Hanus, M. Kinnaert, and J.L. Henrotte. Conditioning technique, a general antiwindup and bumpless transfer method. *Automatica*, 23(6):729–739, 1987.

A. Kestenbaum, R. Shinnar, and F.E. Thau. Design concepts for process control. *Ind. Eng. Chem. Process Des. Dev.*, 15(1):2–13, 1976.

J. Lee and J. Choi. In-line mixer for feedforward control and adaptive feedback control of pH processes. *Chem. Eng. Sci.*, 55:1337–1345, 2000.

J. Lee, S.D. Lee, Y.S. Kwon, and S. Park. Relay feedback method for tuning of nonlinear pH control systems. *AIChE J.*, 39:1093, 1993.

Chapter 5

Frequency Analysis

In Chap. 4, it was shown that to get a system transfer function, a possibility is to subject the system to a sinusoidal input (Fig. 5.1); after a sufficiently long time, the output is a sustained sinusoid of same period, but presents a different amplitude and a phase angle.

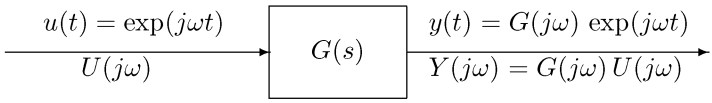


Figure 5.1: Frequency response analysis of a linear system

The study of the characteristics of this sinusoidal output constitutes the analysis of the frequency response of linear processes.

The end of the chapter is dedicated to the introduction of the robustness concept and its use in the synthesis of a controller.

5.1 Response of a Linear System to a Sinusoidal Input

5.1.1 Case of a First-Order Process

To better understand the phenomenon, it is easier to reason about a simple concrete case. Consider a first-order process subjected to a sinusoidal input. The output Laplace transform is related to the Laplace transform of the sinusoidal input $u(t) = A\sin(\omega t)$ by the transfer function

$$Y(s) = G(s)\,U(s) = \frac{K_p}{\tau_p\,s + 1}\,\frac{A\,\omega}{s^2 + \omega^2} \tag{5.1}$$

To know the shape of the time response, $Y(s)$ is decomposed into a sum of fractions that have remarkable denominators; e.g. the following decomposition

can be chosen

$$Y(s) = \frac{a}{\tau_p\, s + 1} + \frac{b\, s + c}{s^2 + \omega^2} \tag{5.2}$$

By use of the values of the poles and identification, the coefficients a, b, c are obtained

$$Y(s) = \frac{K_p\, A}{\tau^2\, \omega^2 + 1}[\frac{\tau^2\, \omega}{\tau s + 1} - \frac{\tau\, \omega\, s}{s^2 + \omega^2} + \frac{\omega}{s^2 + \omega^2}] \tag{5.3}$$

resulting in the output $y(t)$ decomposed in natural response and forced response

$$\begin{aligned} y(t) &= \frac{K_p\, A}{\tau^2\, \omega^2 + 1}[\tau\, \omega\, \exp(-\frac{t}{\tau})] + \frac{K_p\, A}{\tau^2\, \omega^2 + 1}[-\tau\, \omega \cos(\omega\, t) + \sin(\omega\, t)] \\ &= y_n(t) + y_f(t) \end{aligned} \tag{5.4}$$

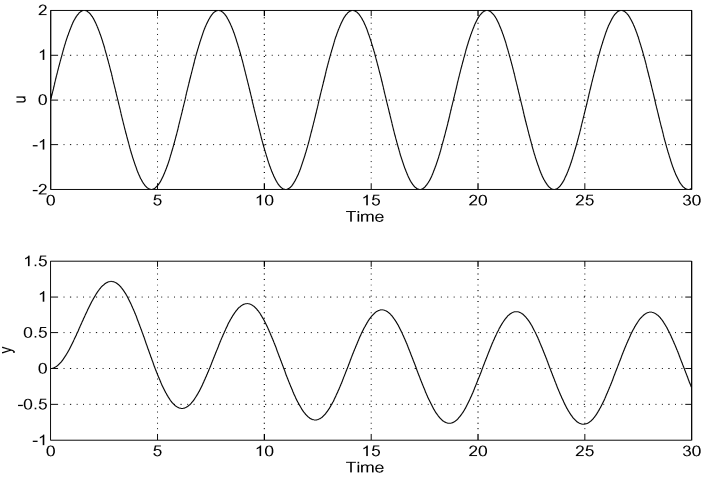


Figure 5.2: Response of a first-order system ($K_p = 2, \tau_p = 5$) to a sinusoidal input ($A = 2, \omega = 1$). Top: input u. Bottom: output y

After a sufficiently long time (Fig. 5.2), the exponential natural response becomes negligible, so that there remains the two sinusoidal terms of the forced response, which can be gathered in the form

$$y(t) = Aa\, \sin(\omega\, t + \phi) \tag{5.5}$$

to take into account the amplitude ratio and the phase angle. Identification gives

$$a \cos(\phi) = \frac{K_p}{\tau^2\, \omega^2 + 1} \quad \text{and} \quad a \sin(\phi) = -\frac{K_p\, \tau\, \omega}{\tau^2\, \omega^2 + 1} \tag{5.6}$$

and

$$a = \frac{K_p}{\sqrt{\tau^2\, \omega^2 + 1}} \quad \text{and} \quad \phi = \text{arctg}(-\tau\, \omega) \tag{5.7}$$

It is important to note that Eq. (5.6) defines ϕ modulo $2k\pi$ while Eq. (5.7) defines ϕ modulo $k\pi$ only. As it corresponds to a physical system, the output is always delayed with respect to the input and the phase angle is defined as a phase lag (see Bode plots).

The response has the same period as the input; the amplitude ratio (coefficient a) and the phase angle depend on pulsation ω (related to the frequency by $\omega = 2\,\pi\,\nu$).

Looking at the process transfer function

$$G(s) = \frac{K_p}{\tau\, s + 1} \tag{5.8}$$

one realizes that:

- The amplitude ratio is equal to the modulus of $G(j\,\omega)$.
- The phase angle is equal to the argument of $G(j\,\omega)$.

This result is, in fact, valid not only in the case of a first-order system, but in the case of any linear system.

5.1.2 Note on Complex Numbers

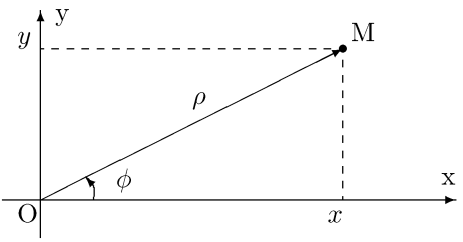


Figure 5.3: Geometric representation of a complex number with abscissa, ordinate, modulus and argument

A complex number z is defined by

$$z = x + jy \qquad \text{with} \quad j^2 = -1 \tag{5.9}$$

where x is the real part and y the imaginary part. The complex number z corresponds to the representation (Fig. 5.3) in the complex plane (Ox, Oy) of vector $\vec{OM}$ (coordinates x, y) where the real part is the abscissa and the imaginary part the ordinate. z can also be expressed in the form

$$\begin{aligned} &z = \rho \exp(j\phi) \qquad \text{with:} \quad \rho = |z| = \sqrt{x^2 + y^2}\ ;\ \phi = (\vec{Ox}, \vec{OM}) \\ &\text{corresponding to:} \quad x = \rho\cos(\phi)\ ;\ y = \rho\sin(\phi) \end{aligned} \tag{5.10}$$

Often, z is given by its modulus and its argument ϕ as

$$|z| = \sqrt{x^2 + y^2}\ ;\ \phi = \text{arctang}(\frac{y}{x}) \tag{5.11}$$

ρ is called the modulus of z and ϕ argument (or angle) of z. Note that Eq. (5.10) defines ϕ modulo $2k\pi$ while Eq. (5.11) defines ϕ modulo $k\pi$ only. Therefore, some information is lost in the passage from Eq. (5.10) to Eq. (5.11).

Let z_1 and z_2 be two complex numbers. The following properties concerning the modules and arguments of products and quotients of complex numbers are important

$$\begin{aligned} |z_1\, z_2| &= |z_1|\,|z_2| \qquad \text{and} \quad \arg(z_1\, z_2) = \phi_1 + \phi_2 \\ \left|\frac{z_1}{z_2}\right| &= \frac{|z_1|}{|z_2|} \qquad \text{and} \quad \arg\left(\frac{z_1}{z_2}\right) = \phi_1 - \phi_2 \end{aligned} \tag{5.12}$$

i.e. the modulus of the product of two complex numbers is equal to the product of their moduli, and the argument of the product is equal to the sum of their arguments. The modulus of the quotient is equal to the ratio of their moduli and the argument of the quotient is equal to the difference of their arguments.

5.1.3 Case of any Linear Process

In a general manner, in open loop, the process output is related to the manipulated input by the process transfer function. In closed loop, the output is related to the set point or to the disturbance by a more complicated global transfer function. In both cases, this transfer function can be described as the ratio of two polynomials such that the numerator degree is always lower or equal to that of the denominator. With both cases being identically treated, only the first case is explained

$$Y(s) = G(s)\,U(s) = \frac{P(s)}{Q(s)}\,U(s) \tag{5.13}$$

The process is subjected to a sinusoidal input $u(t) = A\sin(\omega t)$, hence

$$Y(s) = G(s)\,\frac{A\omega}{s^2+\omega^2} = \frac{P(s)}{Q(s)}\,\frac{A\omega}{s^2+\omega^2} \tag{5.14}$$

The ratio of both polynomials can be decomposed into a sum of rational fractions by using the zeros of $Q(s)$ (poles s_i of $G(s)$), which leads to

$$Y(s) = [\frac{c_1}{s-s_1} + \frac{c_2}{s-s_2} + \ldots] + \frac{a}{s+j\,\omega} + \frac{b}{s-j\,\omega} \tag{5.15}$$

As we are concerned with the behaviour of stable systems, this means that all poles s_i of $G(s)$ are negative real or complex with a negative real part. The output $y(t)$ is composed of the natural response and the forced response. Terms coming from the ratio P/Q correspond to the stable natural response composed of exponential functions with real or complex coefficients, but in all cases have a damped behaviour when t becomes very large. There then remains the two last terms of the forced response, giving the response after a sufficiently long time

$$y(t) = a\,\exp(-j\,\omega\,t) + b\,\exp(j\,\omega\,t) \tag{5.16}$$

To calculate a, it suffices to use the expression of $Y(s)$, multiply by $(s + j\,\omega)$ and set $s = -j\,\omega$, similarly for b, hence

$$a = \frac{A}{-2j} G(-j\,\omega) = \frac{A}{2}\, j\, G(-j\,\omega) \quad \text{and} \quad b = \frac{A}{2j} G(j\,\omega) = -\frac{A}{2} j\, G(j\,\omega) \tag{5.17}$$

The response $y(t)$ is real and it can be checked that a and b are conjugate complex. The response after a sufficiently long time can be written as

$$y(t) = \frac{A}{2}\, j\, [G(-j\,\omega)\, \exp(-j\,\omega\, t) - G(j\,\omega)\, \exp(j\,\omega\, t)] \tag{5.18}$$

and can be simplified by writing that

$$G(j\,\omega) = |G(j\,\omega)|\, \exp(j\,\phi) \tag{5.19}$$

giving the output after a sufficiently long time

$$y(t) = A|G(j\,\omega)|\, \sin(\omega\, t + \phi) \tag{5.20}$$

The amplitude ratio is equal to

$$AR = |G(j\,\omega)| \tag{5.21}$$

and the phase angle

$$\phi = \text{argument}\, (G(j\,\omega)) \tag{5.22}$$

5.1.4 Case of Linear Systems in Series

Consider a global system constituted by n linear subsystems in series (Fig. 5.4). The transfer function of the global system is the product of the individual transfer functions

$$G(s) = G_1(s)\, G_2(s)\, \ldots\, G_n(s) \tag{5.23}$$

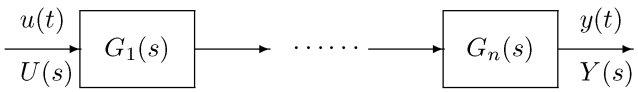


Figure 5.4: System composed of linear subsystems

The modulus $|G(j\,\omega)|$ of the global system transfer function is equal to the product of the moduli of the individual transfer functions

$$|G(j\,\omega)| = |G_1(j\,\omega)|\, |G_2(j\,\omega)|\, \ldots\, |G_n(j\,\omega)| \tag{5.24}$$

while the argument of the global system transfer function is equal to the sum of the arguments of the individual transfer functions

$$\arg\, (G(j\,\omega)) = \arg\, (G_1(j\,\omega)) + \arg\, (G_2(j\,\omega)) + \ldots + \arg\, (G_n(j\,\omega)) \tag{5.25}$$

5.2 Graphical Representation

The influence of frequency on the modulus of $G(j\,\omega)$ and on its argument can be represented by different plots, in particular Bode and Nyquist ones.

5.2.1 Bode Plot

The Bode plot consists of two parts (see example in Fig. 5.5): in the first plot, the logarithm of the modulus of $G(j\,\omega)$ is represented versus the frequency logarithm; in the second plot placed just below, the argument of $G(j\,\omega)$ is represented versus the frequency logarithm.

First-Order System

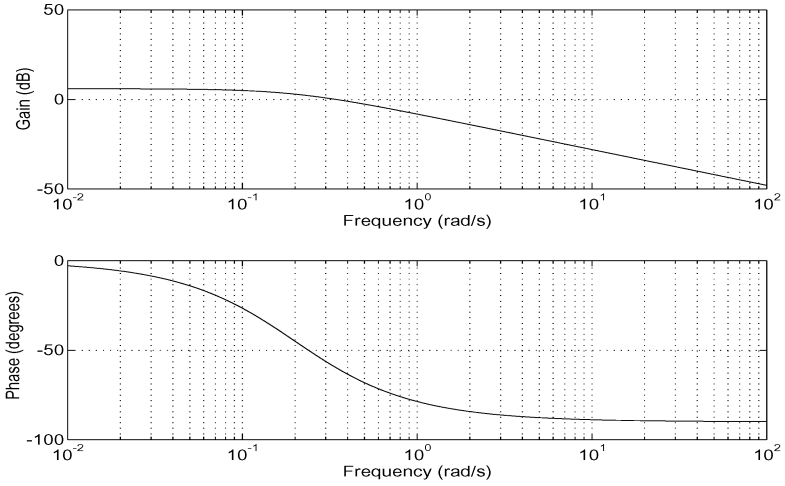


Figure 5.5: Bode plot for a first-order system ($K_p = 2, \tau = 5$)

From the transfer function, one infers

$$G(s) = \frac{K_p}{\tau\, s + 1} \Longrightarrow G(j\,\omega) = \frac{K_p}{\tau\, j\,\omega + 1} \tag{5.26}$$

giving the modulus and argument of $G(j\,\omega)$, respectively

$$|G(j\,\omega)| = \frac{K_p}{\sqrt{\tau^2\,\omega^2 + 1}} \tag{5.27}$$

$$\phi = \text{arctg}(-\tau\,\omega) \tag{5.28}$$

When ω is small (low frequencies), the modulus of $G(j\,\omega)$ is equivalent to the constant process gain K_p (low-frequency asymptote) and the argument ϕ (phase

lag) tends towards 0 (Fig. 5.5). When ω is large (high frequencies), the modulus of $G(j\,\omega)$ is equivalent to $1/\omega$, thus in the log−log plot $[\log(|G(j\,\omega)|), \log(\omega)]$ this gives a straight line of slope -1 (high-frequency asymptote), or, in decibels, -20dB/decade, and the argument (phase lag) tends towards $-\pi/2$.

The modulus expressed in decibels is simply proportional to the logarithm of the modulus expressed without units

$$|G(j\,\omega)|_{\mathrm{dB}} = 20\log(|G(j\,\omega)|) \tag{5.29}$$

Second-Order System

A second-order transfer function is equal to

$$G(s) = \frac{K_p}{\tau^2\,s^2 + 2\,\zeta\,\tau s + 1} \Longrightarrow G(j\,\omega) = \frac{K_p}{1 - \tau^2\,\omega^2 + j\,2\,\zeta\,\tau\,\omega} \tag{5.30}$$

giving the modulus and argument of $G(j\,\omega)$, respectively

$$|G(j\,\omega)| = \frac{K_p}{\sqrt{(1 - \tau^2\,\omega^2)^2\ + (2\,\zeta\,\tau\,\omega)^2}} \tag{5.31}$$

$$\phi = -\mathrm{arctg}(\frac{2\,\zeta\,\tau\,\omega}{1 - \tau^2\,\omega^2}) \tag{5.32}$$

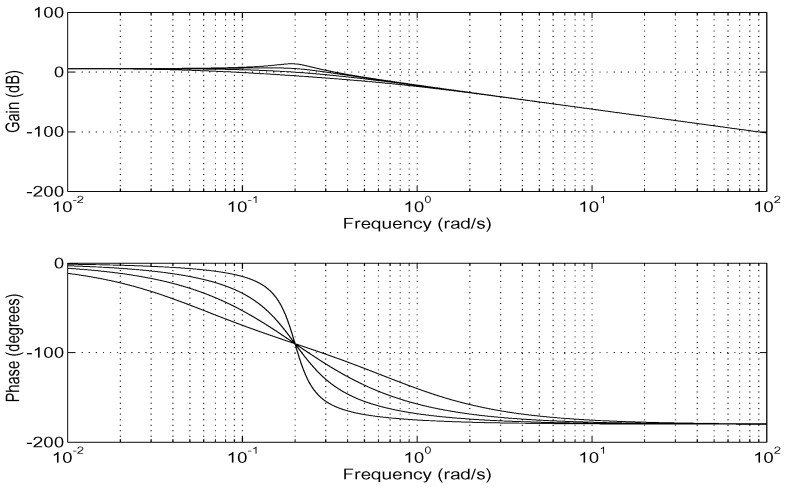


Figure 5.6: Bode plot for a second-order system for different values of ζ ($K_p = 2$, $\tau = 5$, $\zeta = 0.2\,;0.5\,;1\,;2\,;$)

When ω is small (low frequencies), the modulus of $G(j\,\omega)$ is equivalent to the process constant gain K_p and the argument ϕ (phase lag) tends towards 0 (Fig. 5.6). When ω is large (high frequencies), the modulus of $G(j\,\omega)$ is equivalent

to $1/\omega^2$, thus in the log−log plot, it corresponds to a straight line of slope -2 and the argument (phase lag) tends towards $-\pi$.

When a plot is drawn for different values of ζ, the gain curve presents a resonance (amplitude ratio higher than its value on the low-frequency asymptote) for $\zeta = 0.2$ and $\zeta = 0.5$ at the corner (or break) frequency $\omega = \sqrt{1-2\zeta^2}/\tau$, which disappears for $\zeta > \sqrt{2}/2$. To check it, τ being fixed, it is sufficient to search the maximum of $|G(j\omega)|$ with respect to ω. This maximum is equal to $K_p/(2\zeta\sqrt{1-2\zeta^2})$.

5.2.2 n-th-Order System

Consider an n-th-order system with transfer function $G(s)$. At low frequency, the modulus of $G(j\,\omega)$ tends towards the process constant gain K_p and the argument ϕ tends towards 0. At high frequency (ω large), the modulus of $G(j\,\omega)$ is equivalent to $1/\omega^n$, thus in the log−log plot, it corresponds to a straight line of slope $-n$ and the argument tends towards $-n\pi/2$.

Inverse Response System

Inverse response systems are considered as nonminimum-phase systems. To understand this problem, the influences of the numerator and denominator of a transfer function present under the form

$$G(s) = \frac{a\,s+1}{b\,s+1} \tag{5.33}$$

are decomposed, where a is positive.

A term as $(b\,s+1)$ in the transfer function denominator is called a process "lag" (or delay), as the output is delayed with respect to the input: the phase angle ϕ is negative. On the other hand, a term as $(a\,s+1)$ in the transfer function numerator is called process "lead" (or advance), as the output is in advance with respect to the input: the phase angle ϕ is positive. Thus, transfer functions as (5.33) are called "lag−lead".

Suppose that the transfer function is simply

$$a\,s+1 \tag{5.34}$$

It follows that

$$|G(j\omega)| = \sqrt{a^2\omega^2+1} \tag{5.35}$$

$$\phi = \arg(G(j\omega)) = \operatorname{arctg}(a\,\omega) \tag{5.36}$$

The amplitude ratio has a positive slope equal to $+1$ at high frequencies, thus ϕ is included between 0 and $+\pi/2$. This implies that the amplitude ratio tends towards infinity when frequency becomes infinitely large, which has no physical sense. This corresponds to a negative zero for the process.

If the system transfer function is

$$G(s) = \frac{1-a\,s}{b\,s+1}$$

with $a > 0$, the system now presents a positive zero. Then it is an inverse response system. Notice that the amplitude ratio is the same as previously, but that the phase angle corresponding to the numerator is the opposite: $\phi = -\text{arctg}(a\,\omega)$. A positive zero (in the right half-plane) contributes to a phase lag of the global frequency response.

Processes having a zero in the right half-plane or a pure delay are called nonminimum-phase because they present more phase lag than any transfer function that have the same amplitude ratio.

System with Time Delay

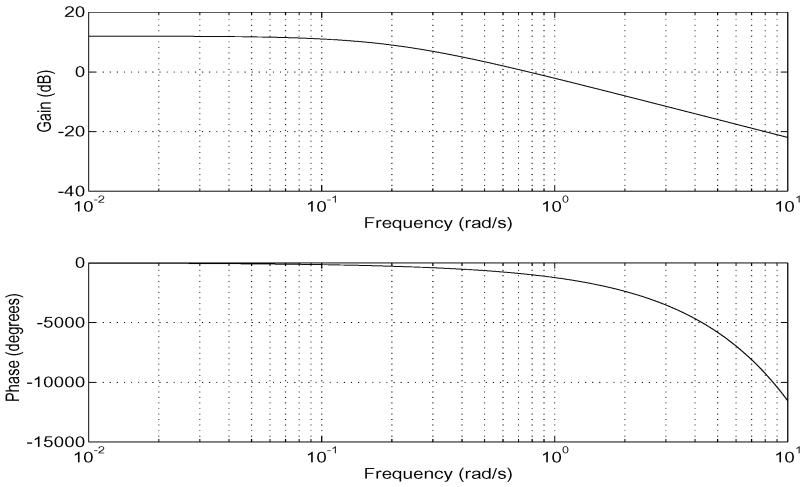


Figure 5.7: Bode plot for a first-order system with time delay ($K_p = 2, \tau_p = 5, t_d = 20, K_c = 2$)

The delay θ is characterized by the presence of a nonlinear term of the form $\exp(-\theta s)$ in the transfer function, which can be written in the form of a product of the exponential term and a transfer function without time delay denoted by G_{wd}

$$G(s) = \exp(-\theta s)\, G_{wd}(s) \tag{5.37}$$

The transfer function for $s = j\omega$ is then equal to

$$G(j\omega) = \exp(-j\omega\theta)\, G_{wd}(j\omega) \tag{5.38}$$

thus its modulus is equal to

$$|G(j\omega)| = |G_{wd}(j\omega)| \tag{5.39}$$

and its argument

$$\arg(G(j\omega)) = \arg(G_{wd}(j\omega)) - \omega\theta \tag{5.40}$$

thus the phase angle is not bounded (Fig. 5.7). This system is nonminimum-phase.

5.2.3 Nyquist Plot

The Nyquist plot (Fig. 5.14) consists of representing in the complex plane, in abscissa, the real part of $G(j\,\omega)$ and, in ordinate, the imaginary part of $G(j\,\omega)$. Thus, the locus is that of a point M such that the vector $O\vec{M}$ corresponds to the complex number $G(j\,\omega)$. It will verify $OM = |G(j\,\omega)|$ and $(\vec{Ox}, O\vec{M}) = \arg(G(j\,\omega))$. The Nyquist plot gives the same information as the Bode plot, but in a different form.

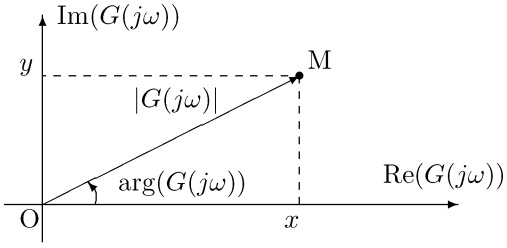


Figure 5.8: Principle of Nyquist representation

First-Order System

The transfer function of a first-order system

$$G(s) = \frac{K_p}{\tau\, s + 1} \tag{5.41}$$

gives

$$\mathrm{Re}[G(j\,\omega)] = \frac{K_p}{\tau^2\,\omega^2 + 1} \tag{5.42}$$

$$\mathrm{Im}[G(j\,\omega)] = \frac{-K_p\,\tau\,\omega}{\tau^2\,\omega^2 + 1} \tag{5.43}$$

Notice that the real part is always positive and the imaginary part always negative, thus the curve is entirely situated in the lower-right quadrant (Fig. 5.9); the angle $(\vec{Ox}, O\vec{M}) = \arg(G(j\,\omega))$ is included between 0 and $-\pi/2$. In this figure, the locus is represented as a full line for positive frequencies and as a dotted line for negative frequencies. When the pulsation ω is zero, the representative point is at K_p on the x axis; then it moves towards the coordinates' origin (limit for ω tending towards infinity). The tangent at the coordinates' origin is vertical. The locus for positive frequencies is, in fact, a half-circle centred on the abscissa axis.

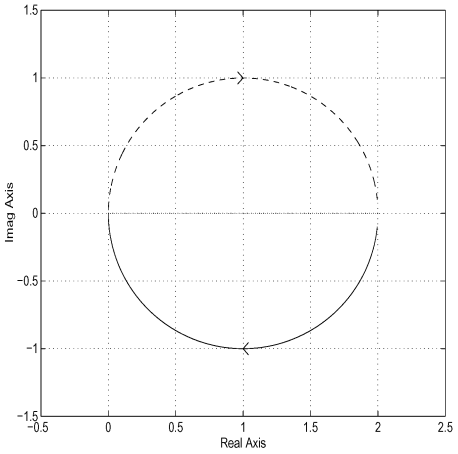


Figure 5.9: Nyquist plot for a first-order system ($K_p = 2, \tau_p = 5$)

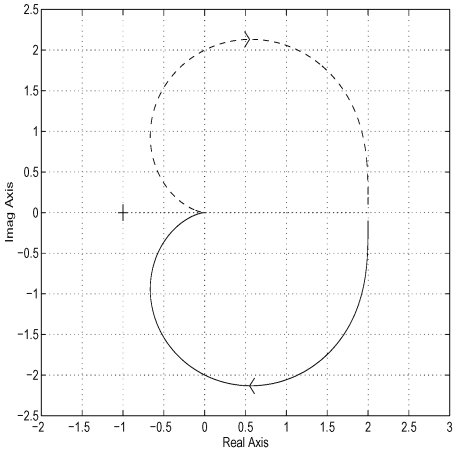


Figure 5.10: Nyquist plot for a second-order system ($K_p = 2\,,\ \tau = 5\,,\ \zeta = 0.5\,$)

Second-Order System

The transfer function of a second-order system

$$G(s) = \frac{K_p}{\tau^2\, s^2 + 2\,\zeta\,\tau\, s + 1} \tag{5.44}$$

gives

$$\mathrm{Re}[G(j\,\omega)] = \frac{K_p(1 - \tau^2\,\omega^2)}{(1 - \tau^2\,\omega^2)^2 \;+\; (2\,\zeta\,\tau\,\omega)^2} \tag{5.45}$$

$$\mathrm{Im}[G(j\,\omega)] = \frac{-K_p\,2\,\zeta\,\tau\,\omega}{(1 - \tau^2\,\omega^2)^2 \;+\; (2\,\zeta\,\tau\,\omega)^2} \tag{5.46}$$

The imaginary part is always negative, thus the curve is located in the lower half-plane (Fig. 5.10); the angle $(\vec{Ox}, \vec{OM}) = \arg(G(j\,\omega))$ is included between 0 and $-\pi$. When the pulsation ω is zero, the point is in K_p on the x axis; then it moves towards the coordinates' origin (limit for ω tending towards infinity). The tangent at the coordinates' origin is horizontal.

5.2.4 n-th-Order System

Consider a general n-th-order system with transfer function $G(s)$. At low frequency, (ω tends towards 0), the representative point M in the Nyquist plane is on the abscissa axis with its abscissa equal to the steady-state gain of the transfer function. When ω increases, the point moves with a decreasing modulus $\rho = OM$ and a negative argument. At high frequency, the point tends towards the origin of the coordinates. Finally, the angle $(\vec{Ox}, \vec{OM})$ describes an angle of $-n\pi/2$ when ω increases from 0 to ∞.

System with Time Delay

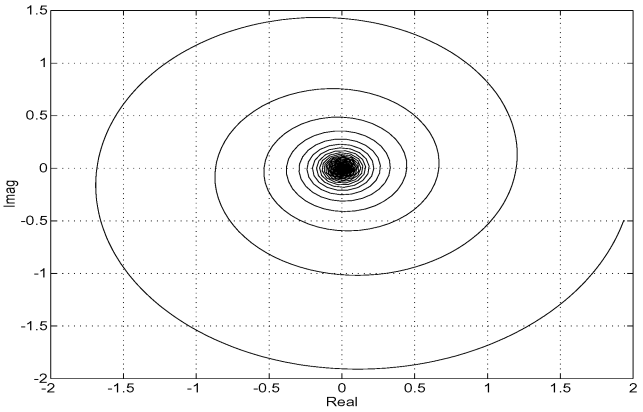


Figure 5.11: Nyquist plot for a first-order system with time delay ($K_p = 2, \tau_p = 5, t_d = 20, K_c = 2$)

Taking again previous notations for the transfer function with time delay

$$G(s) = \exp(-\theta s)\, G_{wd}(s) \Longrightarrow G(j\omega) = \exp(-j\omega\theta)\, G_{wd}(j\omega) \tag{5.47}$$

the real and imaginary parts are deduced

$$\mathrm{Re}[G(j\,\omega)] = \mathrm{Re}[G_{wd}(j\,\omega)]\ \cos(\omega\theta) + \mathrm{Im}[G_{wd}(j\,\omega)]\ \sin(\omega\theta) \tag{5.48}$$

$$\mathrm{Im}[G(j\,\omega)] = \mathrm{Im}[G_{wd}(j\,\omega)]\ \cos(\omega\theta) - \mathrm{Re}[G_{wd}(j\,\omega)]\ \sin(\omega\theta) \tag{5.49}$$

Due to the presence of periodical cosinus and sinus factors in the expressions of the real and imaginary parts of the transfer function with time delay, and also due to the decrease towards zero of the modulus of $G_{wd}(j\,\omega)$ when the frequency increases, the curve in the Nyquist plot for a first-order transfer function with time delay has a spiral aspect. When the frequency tends towards 0, the representative point of maximum modulus of G is far on the right on the abscissa axis, and as soon as frequency increases, the representative point converges towards the origin of the coordinates (Fig. 5.11).

5.2.5 Black Plot

The Black plot of any transfer function is obtained by choosing as abscissa the argument of transfer function $\arg(G(j\omega))$ according to a linear scale and as ordinate the modulus $|G(j\omega)|$ according to a logarithmic scale.

5.3 Characterization of a System by Frequency Analysis

Previously, it was shown that it is possible to characterize linear systems by the analysis of the frequency response to a sinusoidal input, knowing the process transfer function.

In fact, the inverse problem is often set as: given the values of a real process frequency response, what transfer function represents this process? This is an identification question.

The curves or the values of a real process frequency response can be obtained in different ways:

- A direct sinusoidal input: a series of runs with sinusoids of different frequencies; the amplitude ratio and the phase angle are directly determined from input and output signals.
- Impulses which theoretically provide the same information but need a mathematical treatment.
- Random or pseudo-random inputs (pseudo-random binary or ternary sequences) used in advanced identification techniques.

Transfer functions can be deduced from the frequency response in two ways:

- By a simple reasoning on the slopes of the amplitude ratio, the limits of the phase angle (bounded or not).
- By mathematical treatment, e.g. by using a least squares method after transformation of the transfer function.

In the following, it is assumed that the process transfer function is well known. The question is then to design the controller. The first condition that the controller must fulfil is stability. The aim will be to get suitable characteristics of the time response.

5.4 Frequency Response of Feedback Controllers

5.4.1 Proportional Controller

From the controller transfer function

$$G_c(s) = K_c \tag{5.50}$$

it follows that the modulus of $G_c(j\omega)$ is equal to the controller gain K_c and that the phase angle is zero for any frequency.

5.4.2 Proportional-Integral Controller

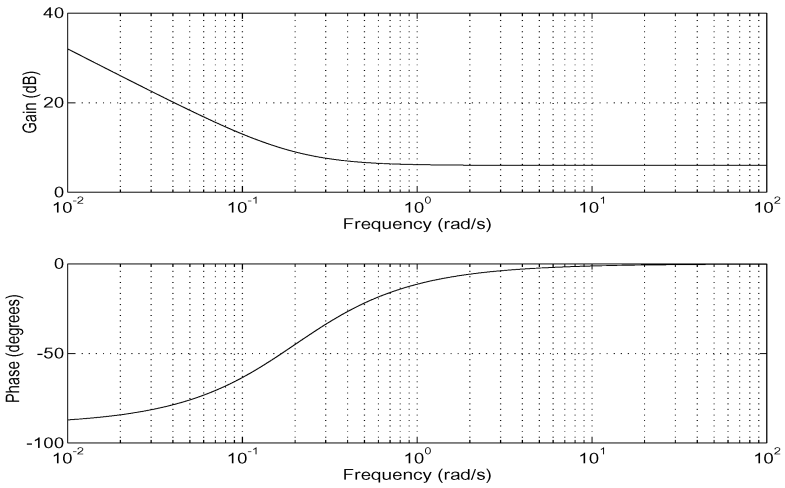


Figure 5.12: Bode plot for a PI controller ($K_c = 2$, $\tau_I = 5$)

The PI controller transfer function is equal to

$$G_c(s) = K_c \left(1 + \frac{1}{\tau_I s}\right) \tag{5.51}$$

Replacing s by $j\omega$, the amplitude ratio results

$$AR = |G_c(j\omega)| = K_c \sqrt{1 + \frac{1}{\tau_I^2 \omega^2}} \tag{5.52}$$

and the phase angle

$$\phi = \arg(G_c(j\omega)) = \text{arctg}(-\frac{1}{\tau_I \omega}) \tag{5.53}$$

At low frequencies (Fig. 5.12), the integral action is dominant, the slope of the straight line of the amplitude ratio is equal to -1 and the phase angle is $-\pi/2$. When the frequency is high, the PI controller behaves as a simple proportional controller, the amplitude ratio tends towards K_c (slope = 0) and the phase angle towards 0.

5.4.3 Ideal Proportional-Derivative Controller

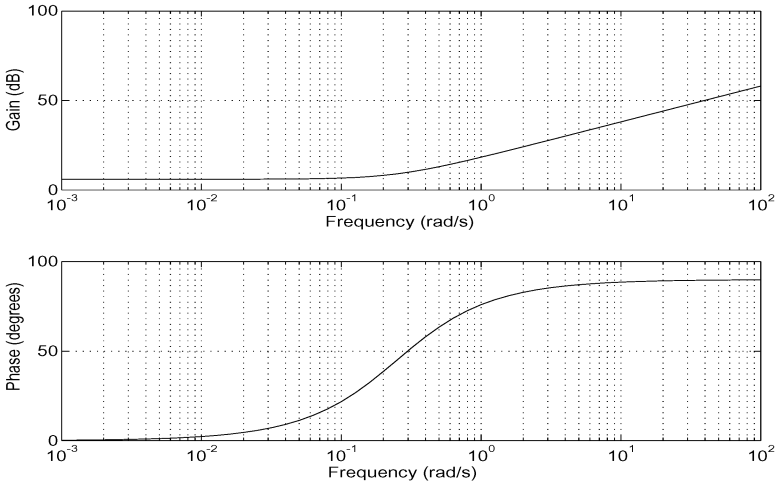


Figure 5.13: Bode plot for an ideal PD controller ($K_c = 2$, $\tau_D = 4$)

Although this type of controller is not physically realizable, it is interesting to look at its Bode plot in order to extract the characteristics of the derivative action. The controller transfer function is equal to

$$G_c(s) = K_c \ (1 + \tau_D s) \tag{5.54}$$

Replacing s by $j\omega$, the amplitude ratio results

$$AR = |G_c(j\omega)| = K_c\sqrt{1 + \tau_D^2\omega^2} \tag{5.55}$$

and the phase angle

$$\phi = \arg(G_c(j\omega)) = \text{arctg}(\tau_D\omega) \tag{5.56}$$

is positive for any value of ω, thus it constitutes an advance.

At low frequencies (Fig. 5.13), the PD controller behaves as a simple proportional controller, the amplitude ratio tends towards K_c (slope = 0) and the phase angle towards 0.

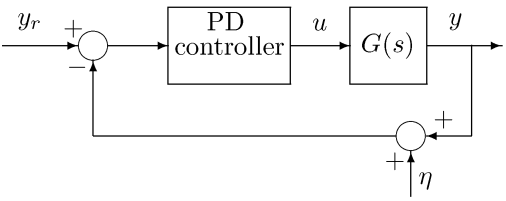


Figure 5.14: Influence of noise on a system controlled by an ideal PD controller

At high frequencies, the derivative action dominates, the slope of the straight line of the amplitude ratio is equal to +1 and the phase angle is $+\pi/2$.
Assume that a sinusoidal measurement noise of the form $\eta = a \sin \omega t$ influences a system controlled by an ideal PD controller (Fig. 5.14). Considering only the pure derivative action of the PD controller, the presence of the noise adds to the input u a signal u_η equal to $u_\eta = -K_c\,\tau_D \dfrac{d\eta}{dt} = -a\,K_c\,\tau_D\,\omega\,\cos(\omega t)$, so that if the noise contains high frequencies, the control signal will have a very large amplitude. To solve this problem, the ideal derivative action is modified by filtering with a first-order filter of time constant τ_D/N so that the real derivative transfer function will be in the form

$$K_c \frac{\tau_D\, s}{\tau_D\, s/N + 1} \tag{5.57}$$

or in a slightly different form

$$K_c \frac{\tau_D\, s + 1}{\beta\, \tau_D\, s + 1} \qquad \text{with} \quad \beta = 1/N \tag{5.58}$$

The derivative action is thus only sensitive at low and mean frequencies. The gain is limited by N at high frequencies and consequently also the measurement noise at high frequencies. Real PID controllers will cut the gain at high frequencies.

The PD controller is seldom used alone, but an integral action is always added. A PD controller is only usable alone when the process is a pure integrating system.

5.4.4 Proportional-Integral-Derivative Controller

Ideal PID Controller

The transfer function of the ideal PID controller is equal to

$$G_c(s) = K_c \left(1 + \frac{1}{\tau_I s} + \tau_D s\right) \tag{5.59}$$

Because of the derivative term, this controller is not physically realizable. By replacing s by $j\omega$, the amplitude ratio results

$$AR = |G(j\omega)| = K_c \sqrt{1 + (\tau_D \omega - \frac{1}{\tau_I \omega})^2} \tag{5.60}$$

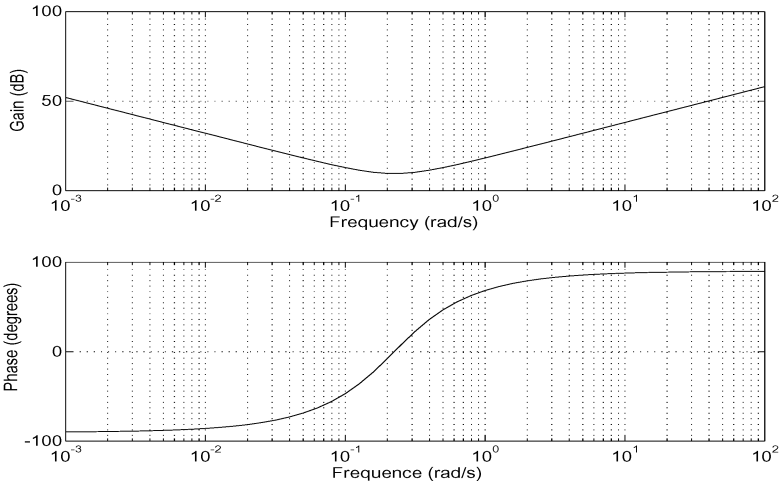


Figure 5.15: Bode plot for an ideal PID controller ($K_c = 2\,,\ \tau_I = 5\,,\ \tau_D = 4$)

and the phase angle

$$\phi = \arg(G(j\omega)) = \text{arctg}(\tau_D\omega - \frac{1}{\tau_I\omega}) \tag{5.61}$$

At low frequencies (Fig. 5.15), the PID controller behaves as a PI controller; the integral action prevails, the amplitude ratio has a slope equal to -1 and the phase angle tends towards $-\pi/2$.

At high frequencies, the derivative action prevails, the slope of the straight line of the amplitude ratio is equal to $+1$ and the phase angle tends towards $+\pi/2$. Thus, the phase angle change is π.

Real PID Controller

With respect to the ideal PID controller, in a real PID controller (Figs. 2.26 and 2.108), the derivative action is modified by filtering by a first-order filter in order to limit the influence of the high-frequency noise. The controller will then be physically realizable.

The transfer function of the real PID controller (pneumatic or electronic) is equal to

$$G_c(s) = K_c \left(\frac{1 + \tau_I\, s}{\tau_I\, s}\right)\left(\frac{\tau_D\, s + 1}{\beta\, \tau_D\, s + 1}\right) \tag{5.62}$$

The values of parameter β are often included between 0 and 1. Its transfer function corresponds, in fact, to the product of the transfer function of a PI and that of a PD.

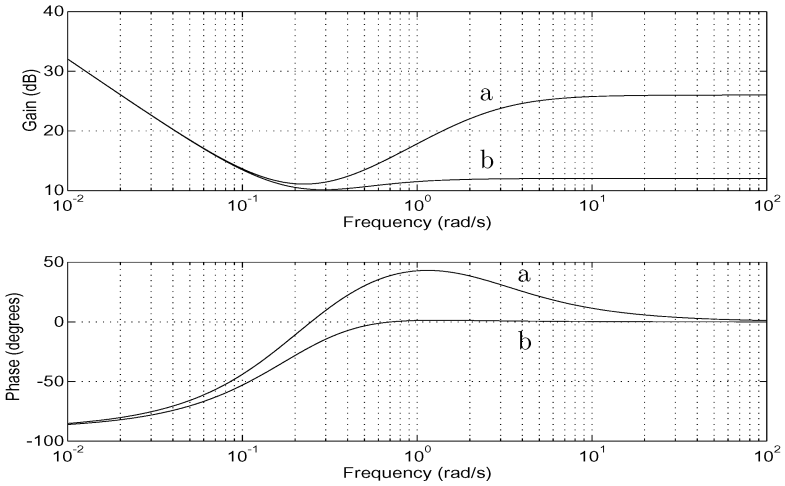


Figure 5.16: Bode plot for a real PID controller ($K_c = 2$, $\tau_I = 5$, $\tau_D = 4$, $\beta = 0.1$ (curves a), and $\beta = 0.5$ (curves b))

After rearrangement, this equation can be written in the form

$$G_c(s) = \frac{1}{\beta\,\tau_D\,s + 1}\,K_c^*\,\left(1 + \frac{1}{\tau_I^*\,s} + \tau_D^*\,s\right) \tag{5.63}$$

The first fraction acts as a first-order filter on a controller that would have its parameters K_c^* , τ_I^* , τ_D^* equal to

$$K_c^* = K_c\left(\frac{\tau_D}{\tau_I} + 1\right) \quad ; \quad \tau_I^* = \tau_D + \tau_I \quad ; \quad \tau_D^* = \frac{\tau_D\,\tau_I}{\tau_D + \tau_I} \tag{5.64}$$

With respect to the ideal PID, the high-frequency behaviour is modified, as the asymptote of the amplitude ratio is then bounded

$$\lim_{\omega\to\infty} AR = \frac{K_c}{\beta} \tag{5.65}$$

For this reason, the controller is less sensitive to the noise. Like in the PD controller, the amplitude ratio is bounded at high frequencies. The derivative action is then limited in the high-frequency domain, which is an advantage as that limits the high-frequency noise influence. Figure 5.16 for $\beta = 0.1$ shows that the amplitude ratio is bounded when ω becomes larger than about 10 rad/s. If larger values of β are chosen, the increase of the amplitude ratio is attenuated and the phase bump disappears ($\beta = 0.5$). For $\beta = 0.9$, the behaviour is very close to a PI controller.

5.5 Bode Stability Criterion

The stability of the closed-loop system is characterized by the solutions of the characteristic Eq. (3.41), which can amount to study of the open-loop transfer function $G_{ol}(s)$. The Bode plot of $G_{ol}(s)$ is a means of realizing this stability study in frequency analysis. Tuning of the PID controller will result.

An important point is that the Bode criterion can only be applied to minimum-phase transfer functions. A system is considered to be minimum-phase if all its poles and zeros are in the left half complex plane. In the opposite case, if at least one pole or zero is in the right half complex plane, the system is said to be nonminimum-phase (Shinners, 1992). In that case, the stability study will be made e.g. by means of the root locus.

This assertion can sometimes be tempered and the Bode criterion can be applied to some nonminimum-phase systems when the reason of that behaviour is known (e.g. the presence of a pure delay).

Consider the first-order process with time delay of Fig. 5.17 for which the following parameters are chosen $K_p = 2$, $\tau = 5$, $t_d = 20$. It is assumed that both transfer functions of the actuator and of the measurement device are equal to 1: $G_a = 1$, $G_m = 1$. As an aside, the controller is set in proportional mode and has gain $K_c = 2$. The process transfer function is equal to

$$G_p(s) = K_p \frac{\exp(-t_d\, s)}{\tau\, s + 1} \tag{5.66}$$

giving the following open-loop transfer function

$$G_{ol}(s) = G_c(s)G_a(s)G_p(s)G_m(s) = K_c K_p \frac{\exp(-t_d\, s)}{\tau\, s + 1} \tag{5.67}$$

Note that $G_{ol}(s)$ is the transfer function of the loop when the latter is considered open. This transfer function is the same as that which plays a key role in the denominator of the closed-loop transfer function and thus of the characteristic equation. The Bode plot of this transfer function shows that because of the exponential term, the phase angle is not bounded and that for a value of the frequency ω_ϕ, called the phase crossover frequency, it is equal to $-\pi$ (Fig. 5.18)

$$\arg\left(G_{ol}(j\omega_\phi)\right) = -\pi \tag{5.68}$$

According to Fig. 5.18 when the frequency is equal to the phase crossover frequency $\omega_\phi \approx 0.13$ rad/s, the amplitude ratio from the Bode plot (gain curve) is equal to around $|G_{ol}(j\omega_\phi)| = (AR)_\phi \approx 3.36$. From the expression of the transfer function, the amplitude ratio is proportional to the gain of the proportional controller, which is taken equal to 2. It results that

$$|G_{ol}| \propto K_c \Rightarrow \frac{|G_{ol}(j\omega_\phi)|}{K_c} \approx \frac{3.36}{2} \approx 1.68 \tag{5.69}$$

In the absence of a Bode plot for the phase angle or of a numerical solver, the phase crossover frequency can be determined by the study of the function

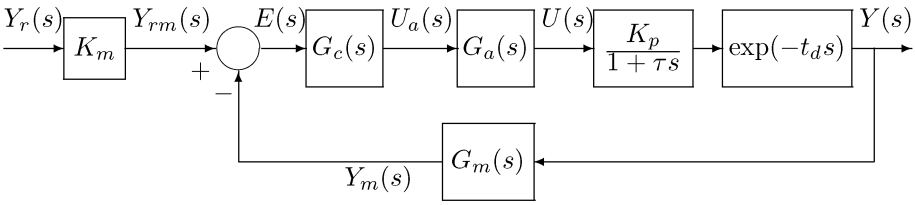


Figure 5.17: First-order system with delay in closed loop

$\arg(G_{ol}(j\omega))$ versus ω, the latter varying in general in the range $[10^{-2}, 10^{+2}]$ rad/time unit for typical chemical processes. If the phase crossover frequency is not found in this range, the latter can be somewhat enlarged. The previous example would thus give Table 5.1.

Table 5.1: Study of the function $\arg(G_{ol}(j\omega))$ versus ω

Frequency (rad/s)	10^{-2}	10^{-1}	0.1285	1	10^{1}	10^{2}
Phase angle (°)	−14.32	−141.15	−180	−1224	−11548	−114681

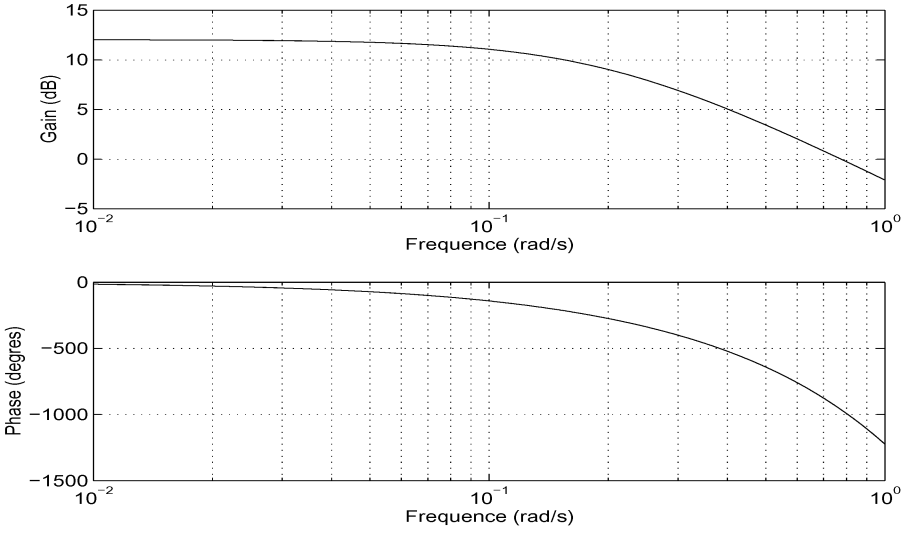


Figure 5.18: Bode plot of the transfer function of the open loop for a first-order system with time delay ($K_c = 2$, $K_p = 2$, $\tau_p = 5$, $t_d = 20$)

The gain of the proportional controller is susceptible to taking any positive value. If the gain K_c is chosen to be equal to the ultimate gain $(K_c)_u$ defined

as

$$(K_c)_u = \frac{K_c}{|G_{ol}(j\omega_\phi)|} \tag{5.70}$$

thus

$$(K_c)_u \approx \frac{2}{3.36} \approx 0.595 \tag{5.71}$$

the amplitude ratio becomes equal to 1 and the measured output y_m is sinusoidal

$$y_m = 1 \sin(0.13\,t - \pi) = -\sin(0.13\,t) \tag{5.72}$$

In fact, the Bode plot is built for the open-loop system response to a sinusoidal set point. By choosing $K_c = (K_c)_u$, the measured output is then exactly in phase opposition with the sinusoidal set point and has the same amplitude.
The gain K_c of the proportional controller, such that the modulus of the open-loop transfer function G_{ol} (or amplitude ratio) is equal to 1, is called the ultimate gain $(K_c)_u$ (Fig. 5.23). Any value of the gain larger than the ultimate gain makes the system unstable.

Consider again the characteristic equation

$$1 + G_{ol} = 0 \tag{5.73}$$

It can be noticed that the first member becomes effectively zero for the phase crossover frequency and the controller gain, such that the amplitude ratio is equal to 1. Thus, one solution of this equation has been found (this equation may possess several of an infinity of solutions for some systems).

Now, consider the following:

- As a response to a sinusoidal set point, the system has been brought into a state of sustained oscillation for a value of the phase crossover frequency; moreover, the system amplitude ratio is equal to 1 for a particular gain of the controller.
- This value of the phase crossover frequency and this gain are kept.
- The loop, which was previously closed, is opened.
- Instantaneously, the set point is set to zero $y_r = 0$. In this case, nothing is changed, as the comparator changes the sign of the measured output, which is then equal to the previous set point. Thus, theoretically, the response must oscillate as a perfect sinusoid.

This point corresponding to the phase crossover frequency ω_ϕ and to the ultimate gain (limit gain) of the controller such that the amplitude ratio is equal to 1 is thus a limit point for the system

- If $K_c > K_{cu} = 0.595$, then the amplitude ratio is larger than 1 for $\phi = -\pi$ and the oscillation of the system will have an increasing amplitude; the system is thus unstable.
- If $K_c < K_{cu} = 0.595$, then the amplitude ratio is lower than 1 for $\phi = -\pi$ and the oscillation of the system will have an decreasing amplitude; the system is thus stable.

Figure 5.19 thus shows this influence of the proportional gain in the neighbourhood of the ultimate gain in the response to a step.

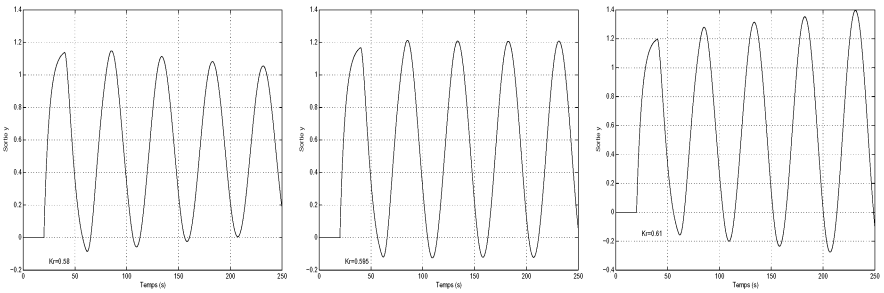


Figure 5.19: Response to a unit-step of the first-order delayed system ($K_p = 2$, $\tau_p = 5$, $t_d = 20$), for three close values of the proportional gain (left $K_r = 0.58$, centre $K_r = 0.595$, right $K_r = 0.61$). Three regimes are displayed from left to right: stability, sustained oscillation, instability

The *Bode stability criterion* results: **a feedback control system is unstable if the amplitude ratio of the corresponding open-loop transfer function is larger than 1 at the phase crossover frequency**.

The Bode stability criterion cannot be applied to systems having an amplitude ratio or a phase angle which does not vary monotonously with respect to frequency.

It may happen that a closed-loop system is unstable for a low value of the gain K_c and that it becomes stable for a larger value of the gain K_c. This is the case of unstable open-loop systems. Consider, for example, the following system

$$G_{ol}(s) = \frac{K_c}{s-2} \tag{5.74}$$

With a positive pole, this system is open-loop unstable. Its phase angle is between $-180°$ and $-90°$ (Fig. 5.20). This system becomes closed-loop stable since $K_c > 2$.

It may also happen that the argument of $|G_{ol}(j\omega)$ takes the value $-\pi$ for more than one frequency ω. In this case, it is necessary to use the Nyquist criterion (Sect. 5.7).

To understand the limitations of Bode criterion, it is recommended to consider the Nyquist plot and the Nyquist criterion. As the Bode plot is easier to obtain manually, it was often preferred with respect to the Nyquist plot. With a computer and a specialized package, this problem disappears.

For the sustained oscillation, which was obtained at the phase crossover frequency for a given value of the controller gain, we must recall the Ziegler-Nichols tuning method of the controller, which was mentioned in Sect. 4.5.2.

For a second-order system regulated by a real PID, the Bode plot of the open-loop transfer function (Fig. 5.21) shows that the phase angle remains between 0 and 180°, thus the system is stable, which can be checked with the closed-loop response to a set point variation.

Note that an open-loop second-order system without delay has a phase angle

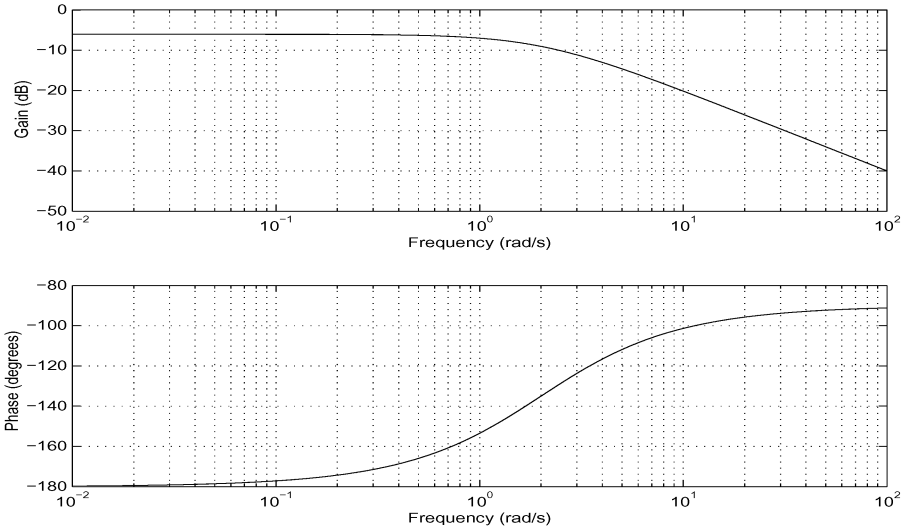


Figure 5.20: Bode plot of the open-loop transfer function of the unstable process (5.74)

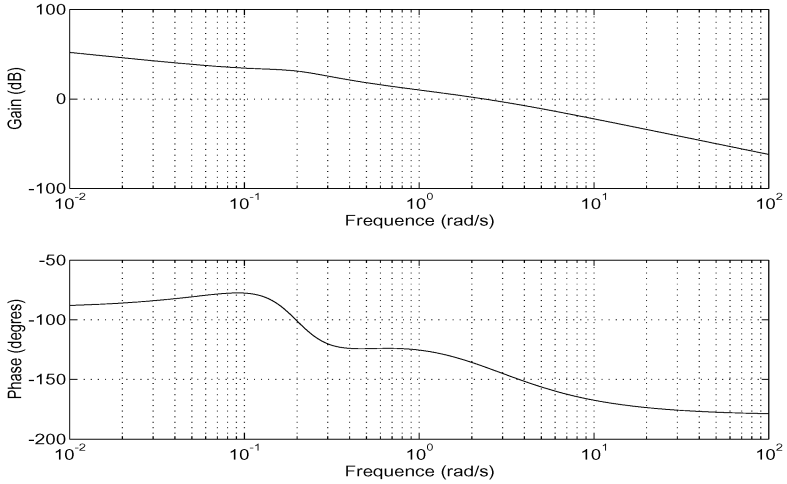


Figure 5.21: Bode plot of the open-loop transfer function for a second-order process ($K_p = 2\,,\, \tau = 5\,,\, \zeta = 0.5$) plus a real PID controller ($K_c = 10\,,\, \tau_I = 5\,,\, \tau_D = 4\,,\, \beta = 0.1$)

always larger than $-\pi$ and thus does not present any phase crossover frequency.

Let the open-loop third-order system

$$G_{ol} = \frac{K_c}{2s^3 + 5s^2 + 7s + 1} \tag{5.75}$$

with $K_c = 10$. From Fig. 5.22, it appears that the phase angle varies between 0 and $-3\pi/2$ and thus this system presents a phase crossover frequency $\omega_\phi \approx 2$ rad/s.

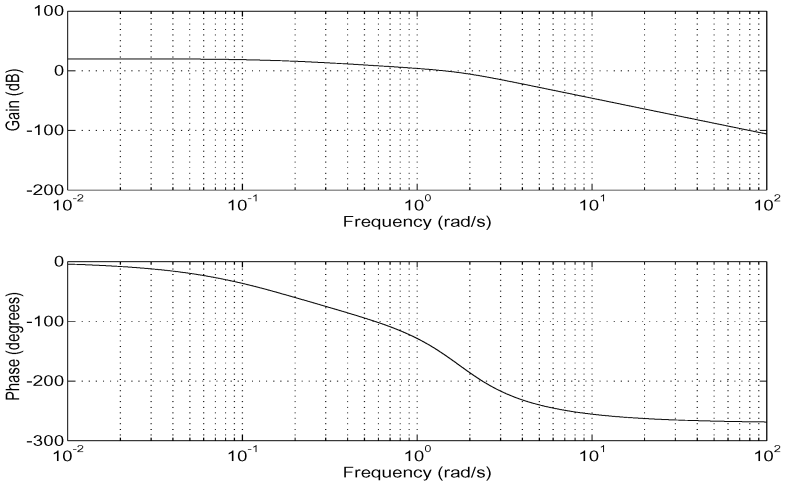


Figure 5.22: Bode plot of the open-loop third-order process transfer function $1/(2s^3 + 5s^2 + 7s + 1)$ plus a proportional controller ($K_c = 10$)

Conclusion: to present a phase crossover frequency, the open-loop transfer function G_{ol} must be at least of third order or present a delay. In the latter case, it can have any order.

5.6 Gain and Phase Margins

The model of a system is always determined with uncertainty. Moreover, the system is susceptible to shift with time. Also, disturbances are present. For these reasons, it is easy to understand that it is unsafe to operate a controller too close to the limit value previously found; a safety margin will have to be provided. To calculate the gain margin and the phase margin, the controller is first set in proportional mode as in the application of the Bode criterion to determine the ultimate gain. This is also related to the tuning by trial and error (Sect. 4.5.1).

5.6.1 Gain Margin

Let AR_ϕ be the amplitude ratio at the phase crossover frequency ω_ϕ (Fig. 5.23); the gain margin is defined as

$$GM = \frac{1}{AR_\phi} = \frac{1}{|G_{ol}(j\omega_\phi)|} \tag{5.76}$$

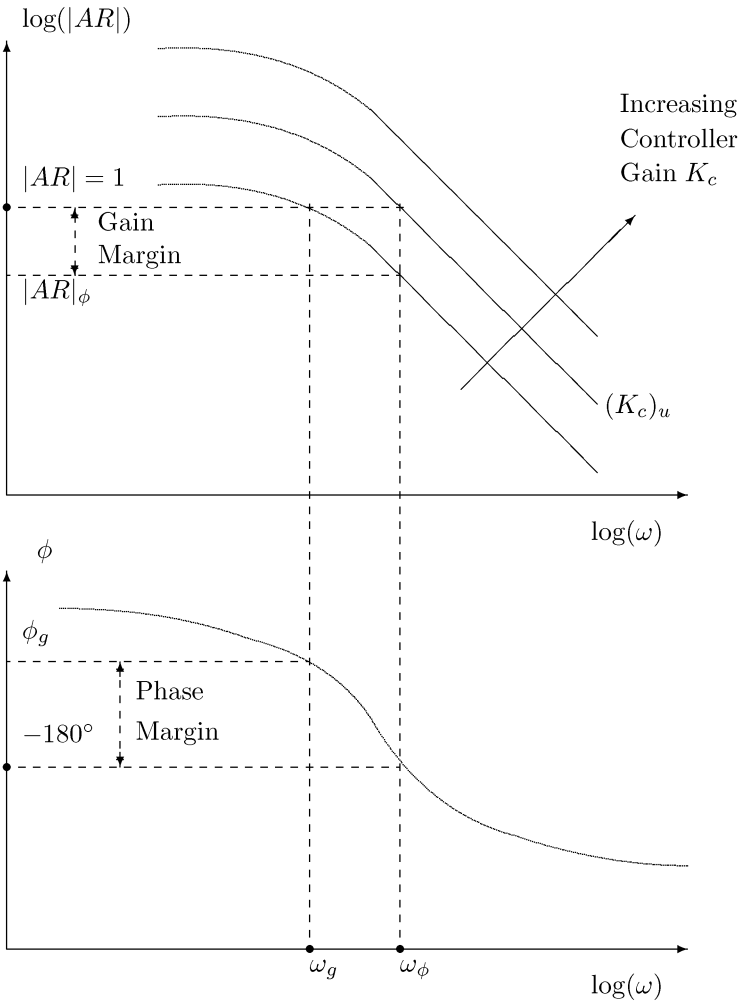


Figure 5.23: Gain and phase margins on Bode plot

If the open-loop transfer function is calculated with the gain of the proportional controller equal to 1, the gain margin is equal to the ultimate gain $(K_c)_u$. Otherwise, it is equal to

$$GM = \frac{(K_c)_u}{K_c} \tag{5.77}$$

where K_c is the gain of the proportional controller.

Frequently, the gain margin is expressed in decibels

$$(GM)_{\text{dB}} = 20\log_{10}(K_c)_u - 20\log_{10}(K_c) \tag{5.78}$$

As the amplitude ratio must be smaller than 1 at the phase crossover frequency, it results that the gain margin must be larger than 1. The larger this value is compared to 1, the further the system is from instability: it becomes robust at the expense of performance.

A recommended value of the gain margin is between 1.7 and 2.

5.6.2 Phase Margin

When the amplitude ratio is equal to 1, the corresponding frequency ω_g is called the gain crossover frequency

$$|G_{ol}(j\omega_g)| = 1 \tag{5.79}$$

and the phase angle is equal to ϕ_g (Fig. 5.23)

$$\phi_g = \arg\left(G_{ol}(j\omega_g)\right) \tag{5.80}$$

If ϕ_g is lower than $-\pi$, the system is unstable; thus this phase angle ϕ_g will have to be larger than $-\pi$. The difference

$$PM = \pi - |\phi_g| \tag{5.81}$$

constitutes the phase margin.

A recommended value of the phase margin is between 30 and 45°.

Recall the tunings recommended by Ziegler and Nichols (Table 5.2) where K_{cu} is the ultimate gain and T_u the ultimate period

$$T_u = 2\pi/\omega_\phi \tag{5.82}$$

corresponding to the sustained sinusoidal oscillation.

Table 5.2: Tunings recommended by Ziegler and Nichols

Controller	K_c	τ_I	τ_D
P	$0.5\,K_{cu}$		
PI	$0.45\,K_{cu}$	$T_u/1,2$	
PID	$0.6\,K_{cu}$	$T_u/2$	$T_u/8$

The value $0.5\,K_{cu}$ corresponds, in fact, to a gain margin of 2, while $0.6\,K_{cu}$ corresponds to a gain margin of 1.7.

When the values recommended by Ziegler-Nichols are taken into account, the PI controller transfer function is equal, at the phase crossover frequency, to

$$G_c(j\,\omega_\phi) = 0.45\,K_{cu}\,(1 - j\frac{1.2}{2\,\pi}) = 0.45\,K_{cu}\,(1 - 0.191j) \tag{5.83}$$

which corresponds to a phase lag of $-10.8°$, while for the PID controller it is equal to

$$G_c(j\,\omega_\phi) = 0.6\,K_{cu}\,(1 - j\frac{1.2}{2\,\pi} + j\frac{2\,\pi}{8}) = 0.6\,K_{cu}\,(1 + 0.467j) \tag{5.84}$$

which corresponds to a phase advance of $25°$.

A reproach often directed at Ziegler-Nichols tuning is that the corresponding controller realizes insufficient damping, so that later a more robust design must be performed.

Example 5.1: **Ziegler-Nichols Tuning**

Consider a very simple example of a first-order system with a time delay ($K_p = 2\,;\ \tau = 10\,;\ t_d = 3$), having the following transfer function

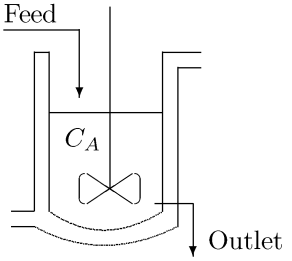
$$G_p(s) = \frac{K_p}{\tau_s + 1}\,\exp(-t_d s) \tag{5.85}$$

and calculate the gain and phase margins for different types of controllers (Table 5.3) tuned by use of the Ziegler-Nichols recommendations. For all the cases, of course, the critical frequency ω_ϕ remains the same. The Bode criterion yields $K_{cu} = 2.945$ and $\omega_\phi = 0.5805$ rad/time unit.

Table 5.3: Gain and phase margins for different types of controllers tuned according to the Ziegler-Nichols recommendations for a given first-order process with time delay

Controller	K_c	τ_I	τ_D	Gain margin	Phase margin (°)	ω_g
P	K_{cu}			1	0	0.5805
P	$0.5\,K_{cu}$			2	62.2	0.2770
PI	$0.45\,K_{cu}$	$T_u/1,2$		2.18	64.2	0.2687
real PID (following)	$0.6\,K_{cu}$	$T_u/2$	$T_u/8$ with $N = 20$	1.47	47.1	0.3439

Example 5.2: **Chemical Reactor PID Control Using Ziegler-Nichols Tuning**



A continuous perfectly stirred chemical reactor with a nonlinear reaction is used to show the performances of the Ziegler-Nichols tuning for a PID. The chemical reactor is simply represented by the following equation

$$\frac{dC_A}{dt} = \frac{F}{V}(C_{A0} - C_A) - k\,C_A^2 \tag{5.86}$$

The objective is to control the reactor concentration by manipulating the feed flow rate. The feed concentration is a disturbance. The parameters and steady-state conditions are given in Table 5.4.

Table 5.4: Parameters and steady state (with superscript ^{s}) conditions of the CSTR

Nominal flow rate of the feed	$F_0^s = 5 \times 10^{-4}$ m^3.s^{-1}
Nominal concentration of reactant A in the feed	$C_{A0}^s = 3000$ mol.m^{-3}
Volume of reactor	$V = 0.1$ m^3
Kinetic constant	$k = 10^{-4}$ m^3.mol^{-1}.s^{-1}
Steady-state concentration of reactant A	$C_A^s = 363.1$ mol.m^{-3}

The volume of the reactor is assumed to be perfectly controlled. The nonlinear model is first linearized, resulting in the following Laplace equation

$$\bar{C}_A(s) = \frac{(C_{A0}^s - C_A^s)\,\dfrac{1}{V}}{s + \dfrac{F^s}{V} + 2\,k\,C_A^s}\,\bar{F}(s) + \frac{\dfrac{F^s}{V}}{s + \dfrac{F^s}{V} + 2\,k\,C_A^s}\,\bar{C}_{A0}(s) \tag{5.87}$$

The actuator transfer and measurement functions are assumed to be, respectively

$$G_a(s) = \frac{0.1}{5\,s + 1} \quad ; \quad G_m(s) = \frac{0.02\,\exp(-5s)}{0.5\,s + 1} \tag{5.88}$$

In the simulation, the process is represented by the nonlinear model while the PID tuning is strictly based on the linearized model: the critical frequency and the ultimate gain used in the Ziegler-Nichols table are drawn from the corresponding Bode plot of the open-loop transfer function.

Three cases are studied:

1. The measurement delay is neglected and a real PID (Fig. 2.27) is used.
2. The measurement delay is neglected and a real PID with anti-windup (Fig. 4.13) is used.
3. The measurement delay $t_d = 5$ s is considered and the real PID of case 1 is used. The results have shown that the improvement by adding an anti-windup is negligible in this particular case.

The parameter N for the real PID is taken as $N = 10$. For the anti-windup, the time constant is $\tau_t = 2$ and the inlet flow rate is assumed to be bound in the range $[0\,,\, 2 \times 10^{-3}]$.

The different characteristics and tunings are gathered in Table 5.5.

Table 5.5: PID tuning by Ziegler-Nichols rule in the absence and in the presence of measurement delay

Measurement delay neglected	
Critical frequency	$\omega_\phi = 0.755$ rad/s
Ultimate gain	$K_{cu} = 0.0602$
PID Tuning by Ziegler-Nichols	
$K_c = 0.0361$; $\tau_I = 4.16$; $\tau_D = 1.04$	

Measurement delay considered	
Critical frequency	$\omega_\phi = 0.206$ rad/s
Ultimate gain	$K_{cu} = 0.0060$
PID Tuning by Ziegler-Nichols	
$K_c = 0.0036$; $\tau_I = 15.26$; $\tau_D = 3.82$	

These characteristics have been implemented and simulated on the "real" process. The process overcomes a step set point from the steady state to $C_A^r = 500$ mol.m^{-3} at time 0, then a step disturbance from the nominal value to $C_{A0} = 3.5 \times 10^3$ mol.m^{-3} at time 150. In the first case (Fig. 5.24), an overshoot corresponding to a saturation of the actuator occurs at the beginning, which disappears when the anti-windup scheme is used in the second case (Fig. 5.25). The overshoot becomes acceptable with use of anti-windup. The disturbance is well rejected. In the third case (Fig. 5.26), the real PID and the same PID with anti-windup give very close results and the overshoot is not very pronounced. However, the influence of the disturbance is more visible in this case of measurement delay and its rejection takes more time.

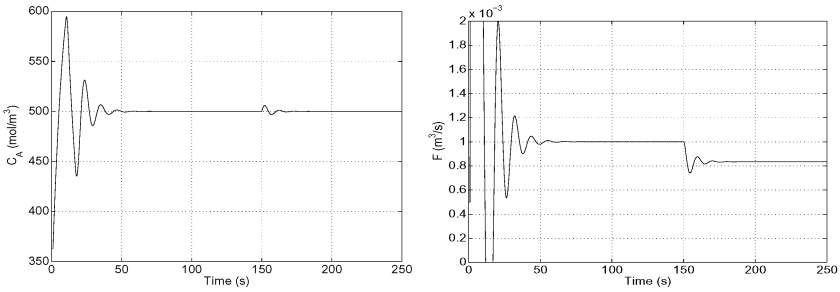


Figure 5.24: Response (left) and input (right) for the closed-loop process without delay with the tuning of the real PID controller obtained by the Ziegler-Nichols rule

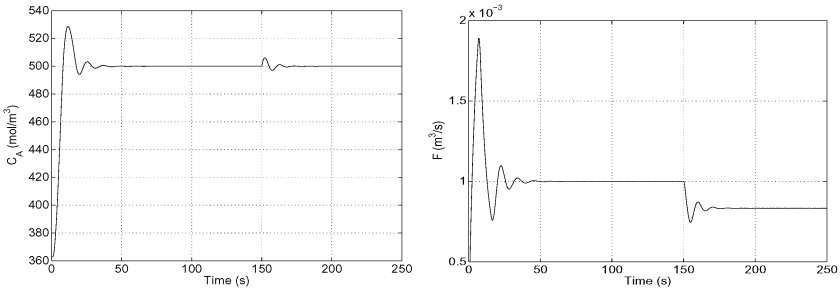


Figure 5.25: Response (left) and input (right) for the closed-loop process without delay with the tuning of the real PID with the anti-windup controller obtained by the Ziegler-Nichols rule

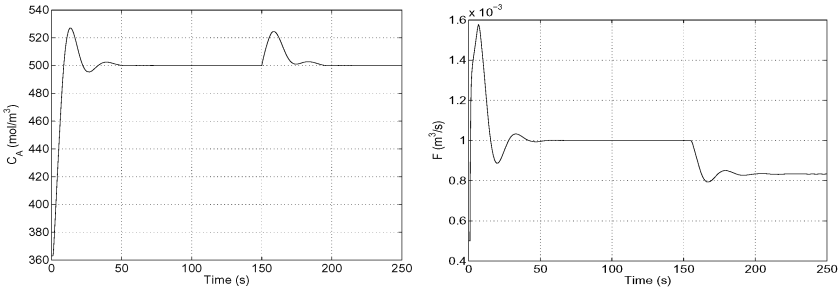


Figure 5.26: Response (left) and input (right) for the closed-loop process with delay with the tuning of the real PID controller obtained by the Ziegler-Nichols rule

5.7 Nyquist Stability Criterion

Recall that the Nyquist plot represents the imaginary part of the open-loop system transfer function with respect to its real part. In the present case

$$G_{ol} = G_c\, G_a\, G_p\, G_m \tag{5.89}$$

thus its modulus is equal to

$$|G_{ol}(j\omega)| = AR = |G_c|\,|G_a|\,|G_p|\,|G_m| \tag{5.90}$$

and its argument

$$\arg(G_{ol}(j\omega)) = \phi_{ol} = \phi_c + \phi_a + \phi_p + \phi_m \tag{5.91}$$

At the phase crossover frequency ω_ϕ and for a gain of the proportional controller K_c such that the amplitude ratio $AR_{ol} = 1$, the open-loop system transfer function G_{ol} verifies

$$1 = |G_c|\,|G_a|\,|G_p|\,|G_m| \tag{5.92}$$

$$-\pi = \phi_c + \phi_a + \phi_p + \phi_m \tag{5.93}$$

In the Nyquist plot, the point of this system is then located at (-1,0). This point constitutes the basis of the Nyquist criterion.

In general, G_{ol} does not present unstable poles (in the right half-plane). The modulus $|G_{ol}|$ is proportional to the controller gain. When the controller gain becomes larger than the ultimate gain, the modulus of the Nyquist vector (representing $G_{ol}(j\omega)$) increases and the curve "turns around" the point $(-1, 0)$ clockwise when the frequency increases; the system is then unstable.

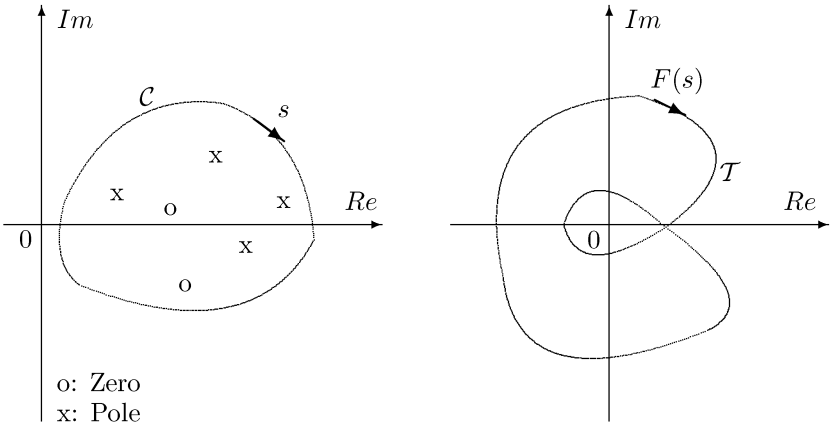


Figure 5.27: Illustration of Cauchy's theorem

The Nyquist criterion relies on the following property of complex functions (Cauchy's theorem).

Consider a function $F(s)$ of complex variable s, a closed contour $\mathcal{C}$ of the complex plane (Fig. 5.27), and assume that $F(s)$ possesses Z zeros and P poles inside the contour $\mathcal{C}$. When s covers the closed contour $\mathcal{C}$, the function $F(s)$ describes a closed trajectory $\mathcal{T}$. The algebraic number N of encirclings (counted positively if they are in the same way as s covers $\mathcal{C}$, negatively in the opposite way) from origin O realized by the function $F(s)$ is equal to the difference $Z - P$:

$$N = Z - P \tag{5.94}$$

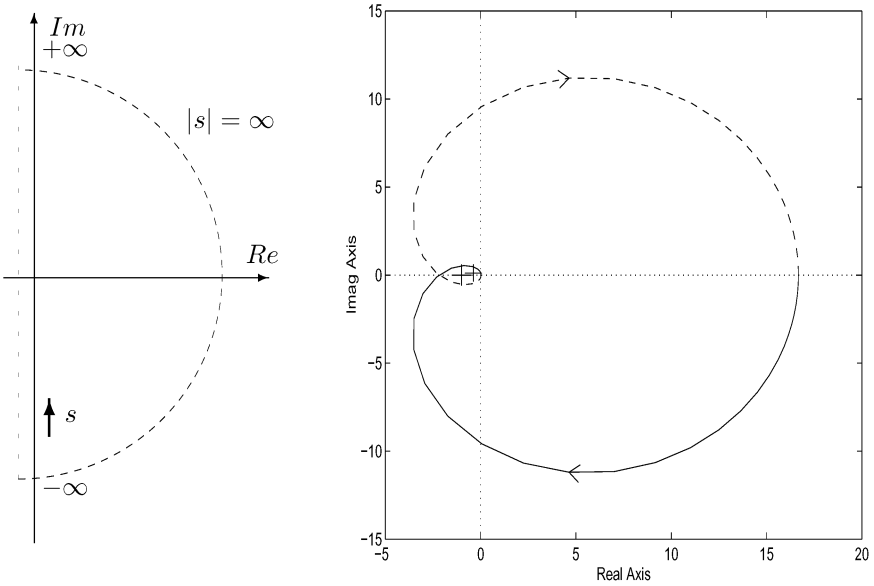


Figure 5.28: Illustration of Nyquist criterion

To get the Nyquist criterion, it suffices to choose the function $F(s) = 1 + G_{ol}(s)$, or by horizontal translation to choose the function $G_{ol}(s)$ with the point $(-1, 0)$ that replaces the origin. The point $(-1, 0)$ is often called the Nyquist point. For the closed-loop transfer function to be stable, $1 + G_{ol}(s)$ must not present positive real part roots, thus $Z = 0$. The contour $\mathcal{C}$, called the Nyquist contour, must then contain all the right half-plane (Fig. 5.28) and is symbolized by a half-circle of infinite radius, centred at the origin, covered clockwise. If one or several of the poles of $G_{ol}(s)$ are present on the imaginary axis, the contour $\mathcal{C}$ turns around them so as not to include them. The algebraic number N of encirclings (counted positively if they are in the same way as s covers $\mathcal{C}$, negatively in the opposite way) of the trajectory $\mathcal{T}$ of $G_{ol}(j\omega)$ around the point $(-1, 0)$ must then be opposite to the number of poles of $G_{ol}(j\omega)$ inside $\mathcal{C}$, thus the right half-plane: $N = -P$.

The *Nyquist criterion* can thus be expressed:

The closed-loop system of transfer function

$$G_{cl} = \frac{G_{ol}}{1 + G_{ol}}$$

is stable if and only if the representative curve of the open-loop system G_{ol} encircles the point $(-1, 0)$, when the *frequency varies from* $-\infty$ *to* $+\infty$, as many times anti-clockwise as G_{ol} possesses poles in the right half-plane (provided there do not exist hidden unstable modes).

When the frequency varies from $-\infty$ to $+\infty$, the locus of G_{ol} is closed. If G_{ol} has no poles in the right half-plane, thus is stable, and if the locus of G_{ol} does not encircle the point $(-1, 0)$, then the closed-loop system is stable. On the contrary, if the locus of G_{ol} encircles the point $(-1, 0)$, then the closed-loop system is unstable.

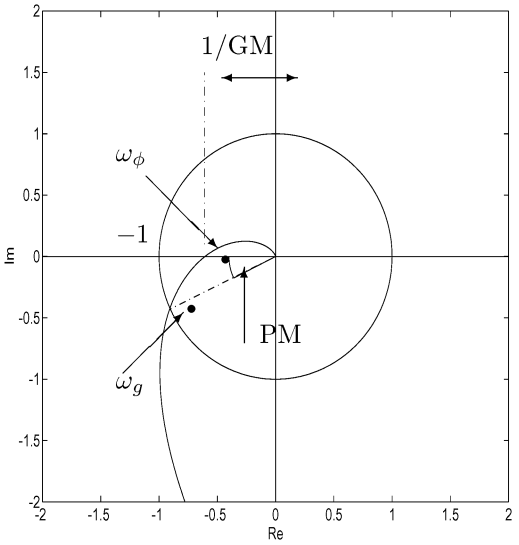


Figure 5.29: Gain and phase margins on the Nyquist plot

The gain margin and the phase margin can be represented on the Nyquist plot, as well as phase and gain crossover frequencies. In the simple case of Fig. 5.29, the gain margin GM is obtained from the intersection of the curve $G_{ol}(j\omega)$ with the real axis, the frequency is then the phase crossover frequency ω_ϕ such that $\arg G_{ol}(j\omega_\phi) = -\pi$. The phase margin is obtained from the intersection of the curve $G_{ol}(j\omega)$ with the circle of unit radius centred at the origin (Fig. 5.29), the frequency is then the gain crossover frequency ω_g such that $|G_{ol}(j\omega_g)| = 1$. The case where there exist several intersections is discussed in Sect. 5.10.

5.8 Closed-Loop Frequency Response

When the closed-loop system is subjected to a set point variation y_r, the output y must present a certain number of characteristics that can be partially

explained (Fig. 5.30) by using the closed-loop frequency response of Y/Y_r.

Let AR_{cl} be the closed-loop amplitude ratio and ψ the phase angle

$$AR_{cl} = |\frac{Y(j\omega)}{Y_r(j\omega)}| \tag{5.95}$$

$$\psi = \arg\left(\frac{Y(j\omega)}{Y_r(j\omega)}\right) \tag{5.96}$$

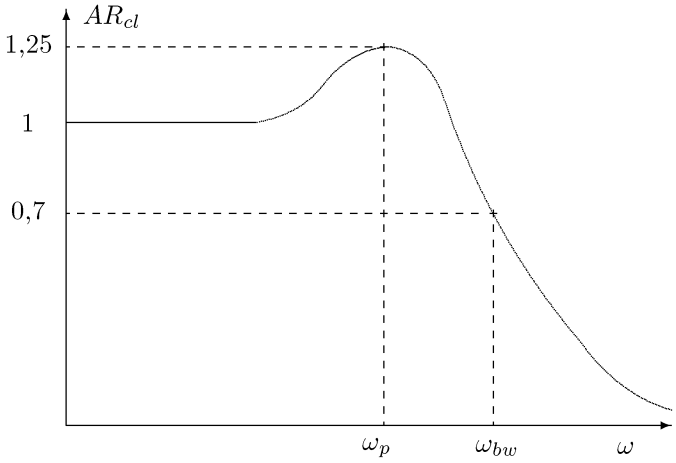


Figure 5.30: Closed-loop desired amplitude ratio for a set point variation

- At low frequencies ($\omega \to 0$), AR_{cl} must be equal to 1, which means that there is no asymptotic deviation of the output with respect to the set point (refer to the final value theorem).
- AR_{cl} must be maintained close to 1 for frequencies as high as possible in order to reach rapidly the new steady state during a set point variation (refer to the initial value theorem).
- It is often desired that the system presents a resonance peak similar to that of a second-order system for which minimum $\zeta = 0.5$. This corresponds to a maximum amplitude ratio at the peak $AR_{cl,p} = 1.25$. The frequency ω_p must be as high as possible; this is in agreement with the previous point.
- The bandwidth ω_{bw} is the frequency for which $AR_{cl} = \sqrt{2}/2 \approx 0.7$. If the bandwidth ω_{bw} is large, the response will be fast with a short rising time.

In a general manner, the ratio Y/Y_r is expressed as

$$\frac{Y(s)}{Y_r(s)} = \frac{G_c(s)G_a(s)G_p(s)K_m}{1 + G_c(s)G_a(s)G_p(s)G_m(s)} \tag{5.97}$$

and in the simple case where the measurement transfer function is a simple gain

$$G_m = K_m \tag{5.98}$$

this ratio Y/Y_r is easily expressed with respect to the open-loop transfer function as

$$\frac{Y(s)}{Y_r(s)} = \frac{G_{ol}(s)}{1 + G_{ol}(s)} \tag{5.99}$$

with $G_{ol}(s) = G_c G_a G_p K_m$

The closed-loop frequency response for set point variations (Fig. 5.31) in the Bode representation depicts the frequency influence on the ratio Y/Y_r. In a logarithmic representation, Fig. 5.30 can be found to again have a slight bump around 2 rad/s.

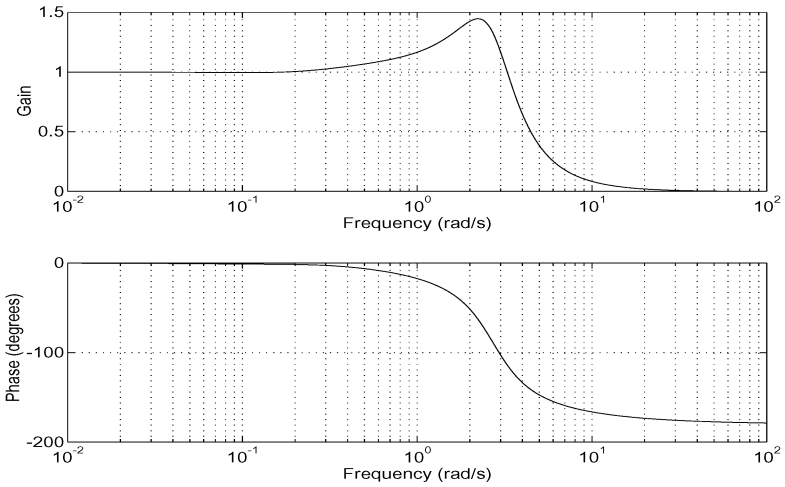


Figure 5.31: Frequency characteristics of the closed-loop second-order process response for set point variations ($K_p = 2\,,\ \tau = 5\,,\ \zeta = 0.5$), plus a real PID controller ($K_c = 10\,,\ \tau_I = 5\,,\ \tau_D = 4\,,\ \beta = 0.1$)

From the previous equation, the closed-loop amplitude ratio results

$$AR_{cl} = \frac{1}{\sqrt{\left[1 + (\cos\phi_{ol}/AR_{ol})\right]^2 + (\sin\phi_{ol}/AR_{ol})^2}} \tag{5.100}$$

and the closed-loop phase angle

$$\psi = \text{arctang}\left(\frac{\sin\phi_{ol}/AR_{ol}}{1 + \cos\phi_{ol}/AR_{ol}}\right) \tag{5.101}$$

where AR_{ol} and ϕ_{ol} are the amplitude ratio and phase angle respectively for the open-loop system transfer function.
The Nichols plot is, in fact, a representation in Black coordinates. For given values of the amplitude ratio and of the phase angle of the open-loop system, it allows us to determine by interpolation the closed-loop values AR_{cl} and ψ. The

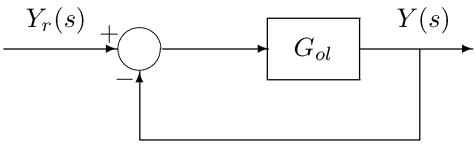
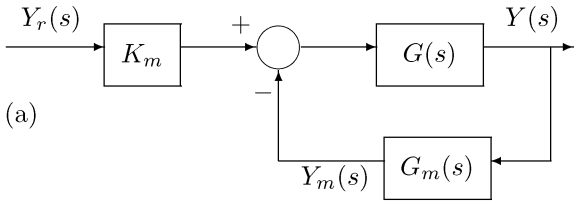


Figure 5.32: Closed-loop unity feedback system

Nichols plot has been built with a unity feedback (Fig. 5.32); this means that the only transfer function of the direct loop is the open-loop process transfer function. In the case where the measurement presents a transfer function (Fig. 5.33), it is necessary to slightly modify this scheme by introducing it into the direct loop and multiplying by the reciprocal of the measurement transfer function before the output (Fig. 5.33). It is possible to check that the closed-loop transfer function is the same for both block diagrams (a) and (b) of Fig. 5.33.

$$Y = \frac{G\,K_m}{1 + G\,G_m}\,Y_r \tag{5.102}$$



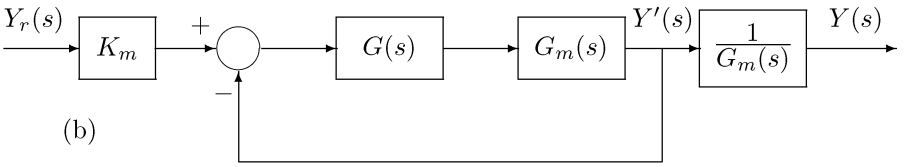


Figure 5.33: Closed-loop system with measurement transfer function (a) and equivalent system with unity feedback (b)

The Nichols plot (Fig. 5.34) considers as abscissa the open-loop phase angle ϕ_{ol} of G_{ol} and as ordinate the open-loop amplitude ratio of G_{ol}. Two families of curves on the graph represent the curves at constant amplitude ratio AR_{cl} (here in dB), and at constant phase angle ψ.

The bold curve is the locus of the points of abscissa ϕ_{ol} and ordinate $AR_{ol} = |G_{ol}|$ when the frequency varies from 0.01 (rectangle at the extremity of the curve) to 100 rad/s for a second-order system controlled by a real PID.

At low frequencies, the bold curve is included between curves −1dB and 1dB; −1dB corresponds to $AR_{cl} = 0.89$ and 1dB corresponds to $AR_{cl} = 1.12$, the excursion for $AR_{cl,\mathrm{dB}} < 0$ being very limited. The recommendations

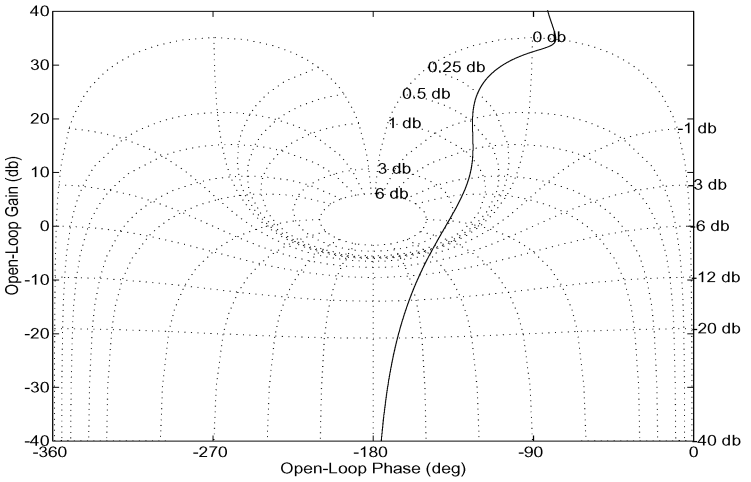


Figure 5.34: Nichols plot of a second-order system ($K_p = 2\,,\ \tau = 5\,,\ \zeta = 0.5$) and a real PID controller ($K_c = 10\,,\ \tau_I = 5\,,\ \tau_D = 4\,,\ \beta = 0.1$)

$1 < AR_{cl} < 1.25$ are thus relatively well respected. Notice that the maximum of AR_{cl} would be denoted by $\|T\|_\infty$ in a robustness study (cf. Sect. 5.10).

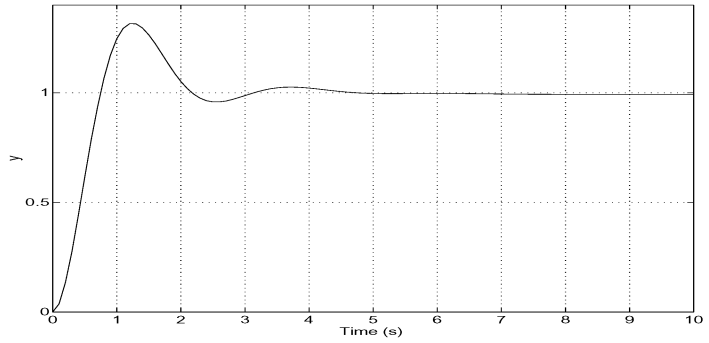


Figure 5.35: Response to a set point step of a second-order system ($K_p = 2\,,\ \tau = 5\,,\ \zeta = 0.5$) and a real PID controller ($K_c = 10\,,\ \tau_I = 5\,,\ \tau_D = 4\,,\ \beta = 0.1$)

Then the response to a set point variation is studied. The response to a set point step of this second-order system controlled by a real PID (Fig. 5.35) presents suitable characteristics with respect to the rising time and the asymptotic deviation. The overshoot may nevertheless seem too large.

What has been done for the set point can also be done for the disturbance;

it is, however, necessary frequently to verify that an optimal tuning for the set point is no more optimal for the disturbance and vice versa.

Again, in the case of the second-order process controlled by a real PID, the ratio Y/D was studied showing the influence of the disturbance. It was assumed that the transfer function G_d between disturbance and output is equal to that of the process G_p. The plot of the closed-loop frequency response for disturbance variations was thus drawn (Fig. 5.36). It is noticed that the gain remains always lower than 1; the aim $|Y/D| \ll 1$ must be reached in the largest possible frequency range.

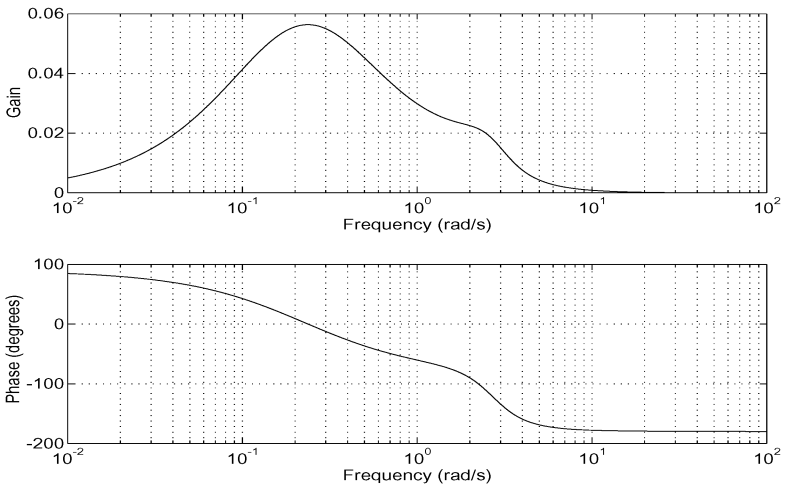


Figure 5.36: Closed-loop frequency response for disturbance variations of a second-order process ($K_p = 2\,,\, \tau = 5\,,\, \zeta = 0.5$) plus a real PID controller ($K_c = 10\,,\, \tau_I = 5\,,\, \tau_D = 4\,,\, \beta = 0.1$)

5.9 Internal Model Principle

The internal model principle is a recommendation of general interest, usable for any controller synthesis.

According to the diagram of Fig. 5.37, the output is equal to

$$Y(s) = \frac{G(s)\,G_c(s)}{1 + G(s)\,G_c(s)}\,Y_r(s) + \frac{1}{1 + G(s)\,G_c(s)}\,D(s) \tag{5.103}$$

The closed-loop frequency response is thus

$$Y(j\omega) = \frac{G(j\omega)\,G_c(j\omega)}{1 + G(j\omega)\,G_c(j\omega)}\,Y_r(j\omega) + \frac{1}{1 + G(j\omega)\,G_c(j\omega)}\,D(j\omega) \tag{5.104}$$

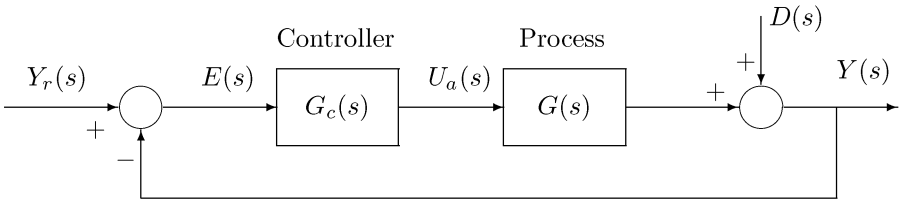


Figure 5.37: Feedback control with disturbance

So that the output correctly follows the desired output, it is sufficient that the open-loop gain $G(j\omega)\,G_c(j\omega)$ is sufficiently large in the dominant frequency domain of the set point $Y_r(j\omega)$ and the disturbance $D(j\omega)$. According to the internal model principle by Francis and Wonham (1976), it is sufficient to place the modes corresponding to these frequencies as well as to high frequencies in the polynomial denominator of the controller $G_c(s)$. If the set point or the disturbance is constant, the dominant frequency is equal to zero, and an integrator $1/s$ must be introduced in the controller, thus s in its denominator. In a general manner, Wohnam (1985) mentions that a controller is structurally stable only if the controller uses the feedback of the controlled variable, and incorporates in the feedback loop a correctly duplicated model of the dynamic structure of the exogenous signals (of external source) that the controller must treat. As an image form, (Wohnam, 1985) says that any good controller must incorporate a model of the external world.

5.10 Robustness

The notion of robustness is related to uncertainties dealing either with the process itself or the process environment, such as unmodelled disturbances or ill-considered process dynamics, frequently at high frequency, in the case where a control model is used. If possible, a control system must be not very sensitive to these factors. However, it must be noted that a compromise performance-robustness must be searched. Clearly, robustness is essential in an industrial context (Larminat, 1991).

A possibility (Morari and Zafiriou, 1989) is to quantify the relative error on the model by

$$e_m = \frac{G - \tilde{G}}{\tilde{G}} \tag{5.105}$$

where G is the actual process model (unknown by definition) and $\tilde{G}$ the used available model. So that the controlled process is stable, it is necessary that

$$\max_{\omega} |e_m| < \frac{1}{AR_{cl,p}} \tag{5.106}$$

$AR_{cl,p}$ being the value of AR_{cl} at the resonance peak.

To get a given value of $AR_{cl,p}$ for the controller, the gain and phase margins GM and PM must satisfy

$$GM \geq 1 + \frac{1}{AR_{cl,p}} \tag{5.107}$$

$$PM \geq 2 \arcsin(1/2AR_{cl,p}) \tag{5.108}$$

Thus, to get $AR_{cl,p} = 1.25$ as previously recommended, this gives $GM = 1.8$ and $PM = 0.823\,\text{rad} = 47°$.

Actually, there exist more formal approaches, such as the H_∞ theory, based on the use of norm and a frequency study (Doyle et al., 1992; Kwakernaak, 1993; Morari and Zafiriou, 1989; Oustaloup, 1994). A brief overview will be given in this section, in order to show all the interest in this theory and to open new horizons.

First, define the usual norms for signals:
2-norm or Euclidean norm of the signal $x(t)$ (of finite energy)

$$\|x\|_2 = \left(\int_{-\infty}^{\infty} x(t)^2 \, dt \right)^{1/2} \tag{5.109}$$

∞-norm of the signal $x(t)$

$$\|x\|_\infty = \max_t |x(t)| \tag{5.110}$$

For a signal $x(t)$ of finite mean power, the square root of the power is defined

$$\lim_{T \to \infty} \left(\frac{1}{2T} \int_{-T}^{T} x(t)^2 \, dt \right)^{1/2} \tag{5.111}$$

Then, similarly, the norms are defined for a system transfer function (with respect to frequency ω)
2-norm of $G(s)$

$$\|G\|_2 = \left(\frac{1}{2T} \int_{-\infty}^{\infty} |G(j\,\omega)|^2 \, d\omega \right)^{1/2} \tag{5.112}$$

∞-norm of $G(s)$

$$\|G\|_\infty = \max_\omega |G(j\,\omega)| \tag{5.113}$$

Thus, the infinity norm of G: $\|G\|_\infty$ represents the maximum amplitude of G in the Bode plot.
Consider the general block diagram (Fig. 5.38) for robustness study. d_2 is a measured disturbance influencing the output y, so that a feedforward G_{ff} is used (cf. Sect. 6.6.3). d_1 is an unmeasured disturbance and η represents a noise that influences measurement. $F(s)$ is a prefilter modifying the set point.

From the open-loop transfer function $G_{ol} = C\,P$, two new functions are defined which allow us to characterize the system:

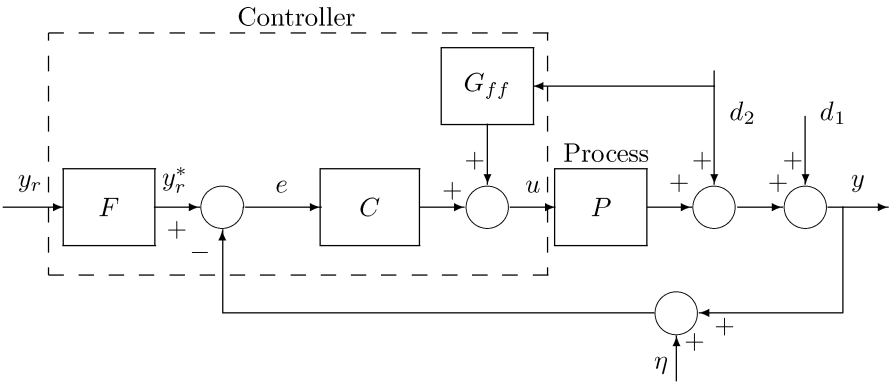


Figure 5.38: Closed loop system considered for robustness study

- The sensitivity function S

$$S(s) = \frac{E(s)}{Y_r^*(s)} = \frac{1}{1 + G_{ol}(s)} \tag{5.114}$$

- The complementary sensitivity function T

$$T(s) = \frac{Y(s)}{Y_r^*(s)} = \frac{G_{ol}(s)}{1 + G_{ol}(s)} \tag{5.115}$$

which is related to the closed-loop transfer function by

$$G_{cl}(s) = \frac{Y(s)}{Y_r(s)} = T(s)\,F(s) \tag{5.116}$$

Assuming that the disturbance d_2 and the associated feedforward controller do not exist, notice that the sensitivity function would also be equal to

$$S(s) = -\frac{E(s)}{D_1(s)} \tag{5.117}$$

and thus represents the sensitivity of the system to a disturbance d_1, as e represents the error. The comparison of both expressions of the sensitivity function shows that the set point and the disturbance act out the same role.

The sensitivity and complementary sensitivity functions depend only on the transfer functions inside the loop and always verify

$$S(s) + T(s) = 1 \tag{5.118}$$

From scheme 5.38, the output is equal to

$$Y(s) = T(s)\,F(s)\,Y_r(s) + S(s)\,D_1(s) + [1 + G_{ff}(s)\,P(s)]\,S(s)\,D_2(s) - T(s)\,N(s) \tag{5.119}$$

It clearly appears, from this equation, that if the influence of the disturbance is to be limited, $S(s)$ must be small and if the influence of the measurement

noise is to be limited, $T(s)$ must be small. A compromise must then be found between the disturbance attenuation and the filtering of measurement errors.

The following relation

$$\lim_{\Delta P \to 0} \frac{\Delta T/T}{\Delta P/P} = \frac{dT}{dP}\frac{P}{T} \tag{5.120}$$

which presents a limited mathematical interest, is rich at the physical level. Indeed, the left member represents the ratio of a relative variation of the complementary sensitivity function T over a relative variation of transfer function P (uncertainty in the process). On the other hand, using the definitions of S and T, it can be shown that the right member is equal to the sensitivity function S. The conclusion is that the sensitivity function measures the sensitivity of the closed-loop transfer function to a process variation.

A criterion of tracking performance[1] could be to postulate that the user's wish is

$$\|S\|_\infty < \epsilon \tag{5.121}$$

in order to minimize the tracking error, the amplitude of which would be lower than ϵ for a set point of amplitude lower or equal to 1.

To limit the influence of noise η, we should have[2]

$$\|T\|_\infty < \epsilon \tag{5.122}$$

The same condition is obtained to limit the influence of the disturbance. Clearly, the objectives of tracking and disturbance rejection on one hand, of noise suppression on the other are contradictory.

It is also possible to filter the set point, e.g. by a lowpass or bandpass filter of transmittance W_1 so that the previous equation would become

$$\|W_1\, S\|_\infty < 1 \tag{5.123}$$

where W_1 represents a frequency weighting function. Equation (5.123) is the condition of nominal performance. Notice that this equation can also be written in the form

$$|S(j\,\omega)| < \frac{1}{|W_1(j\,\omega)|} \quad , \quad \forall\omega \tag{5.124}$$

which means that, in the Bode plot, the sensitivity curve is located below a prescribed curve for any frequency.

---

[1] In the multivariable case, we impose

$$\sigma_{\max}|S(j\omega)| << 1$$

where $\sigma_{\max}$ is the largest singular value of $S(j\omega)$.

[2] In the multivariable case, we would have

$$\sigma_{\max}|T(j\omega)| << 1$$

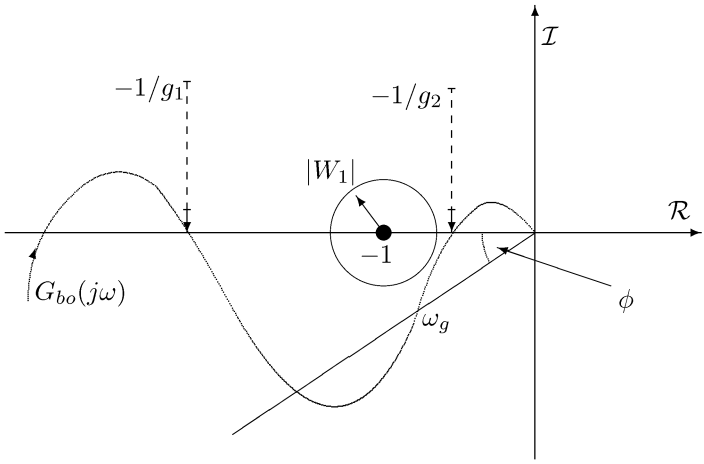


Figure 5.39: Nyquist locus for the open-loop transfer function with the gain margin and the phase margin

From the definition of S, the previous inequality can also be transformed as

$$\left|\frac{W_1(j\,\omega)}{1+G_{ol}(j\,\omega)}\right| \leq 1 \quad , \quad \forall\omega \quad \Longleftrightarrow \quad |W_1(j\,\omega)| < |1+G_{ol}(j\,\omega)| \quad , \quad \forall\omega \tag{5.125}$$

thus, in the Nyquist plot, the locus of $G_{ol}(j\,\omega)$ is situated entirely outside the disc of centre -1, radius $|W_1(j\,\omega)|$ (Fig. 5.39), which corresponds to a more restrictive vision of the traditional Nyquist criterion. In this figure, the gain margin can be defined as the open interval $]g_1, g_2[$ which can be interpreted in the following manner by denoting by $\tilde{P}$ the model of the process P which thus represents an approximation of P, for any g taken in the interval $]g_1, g_2[$, the controller of gain C stabilizes the process P when P is equal to $g\,\tilde{P}$.

Taking the intersection of the locus of $G_{ol}(j\,\omega)$ and of the circle of centre O and unit radius, the point of the gain crossover frequency ω_g is obtained such that $|G_{ol}(j\,\omega_g)| = 1$, and the angle thus formed with the abscissa axis is the phase margin ϕ. If there exist several intersections, the phase margin is the minimum of the formed angles.

The delay margin is defined from the gain margin and the gain crossover frequency ω_g as

$$M_d = \frac{\phi}{\omega_g} \tag{5.126}$$

The delay margin corresponds to the largest delay θ such that the controller of gain C stabilizes the process P when P is of the form $\exp(-\theta s)\,\tilde{P}$.

In fact, these margins can reveal themselves to be insufficient, and the

distance from point -1 to the locus of G_{ol} equal to

$$\frac{1}{\|S\|_\infty} \tag{5.127}$$

is considered. It means that a model error, which was not taken into account by the gain and phase margins must be larger than a certain value so that the process is unstabilized with the chosen controller of gain C.

Now introduce the model error such that the nominal transfer function is P and the modified or perturbed function is in the form

$$\tilde{P} = (1 + \Delta W_2)\, P \tag{5.128}$$

thus ΔW_2 represents the relative error on the process, with Δ being a stable transfer function such that its norm $\|\Delta\|_\infty < 1$, and W_2 is also a stable transfer function, which then verifies

$$\left| \frac{\tilde{P}(j\,\omega)}{P(j\,\omega)} - 1 \right| \leq |W_2(j\,\omega)| \quad , \quad \forall \omega \tag{5.129}$$

Hence, $|W_2(j\,\omega)|$ defines an uncertainty profile such that the point $\tilde{P}/P$ is in a disc of centre 1, radius $|W_2|$. Note that this radius corresponding to uncertainty increases in general with frequency ω.

The robust stability is defined as the property that the closed-loop system is stable with a given controller C when the process P is situated in a ball $\mathcal{P}$.

Then consider the open-loop perturbed function

$$\tilde{P}\,C = (1 + \Delta W_2)\, P\,C = (1 + \Delta W_2)\, G_{ol} \tag{5.130}$$

which will be studied in the Nyquist plane. This function must not encircle the point of abscissa -1, thus consider the quantity

$$1 + \tilde{P}\,C = 1 + (1 + \Delta W_2)\, G_{ol} = (1 + G_{ol})\,(1 + \Delta\, W_2\, T) \tag{5.131}$$

The condition of robust stability for multiplicative uncertainty can be formulated as

$$\|W_2\, T\|_\infty \leq 1 \tag{5.132}$$

and the stability margin is the reciprocal $1/\|W_2\, T\|_\infty$. It results that

$$\|\Delta\, W_2\, T\|_\infty \leq \|W_2\, T\|_\infty \leq 1 \tag{5.133}$$

Equation (5.132) is written again

$$\left| \frac{W_2(j\omega)\, G_{ol}(j\omega)}{1 + G_{ol}(j\omega)} \right| < 1 \quad \forall \omega \quad \Longleftrightarrow \quad |W_2(j\omega)\, G_{ol}(j\omega)| < |1 + G_{ol}(j\omega)| \quad \forall \omega \tag{5.134}$$

which can be expressed graphically (Fig. 5.40) by the fact that the critical point -1 is always located outside the uncertainty discs of centre $G_{ol}(j\omega)$ and radius $|W_2(j\omega)\, G_{ol}(j\omega)|$.

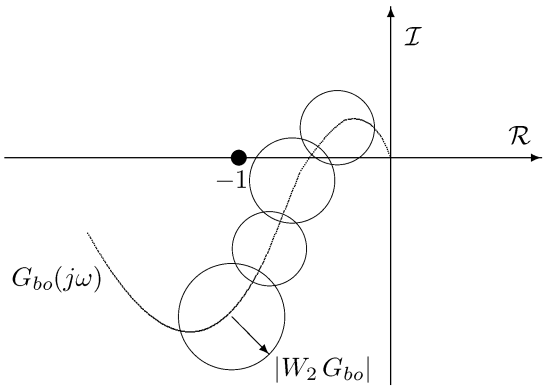


Figure 5.40: Nyquist locus for the open-loop transfer function and uncertainty discs

A system should simultaneously present a good performance and good stability. For this reason, the conditions of the equation of nominal performance (5.123) and of the equation of robust stability (5.132) are gathered under the condition of robust performance

$$\| \, |W_1\,S| + |W_2\,T| \, \|_\infty < 1 \tag{5.135}$$

which allows us to simultaneously guarantee both previous conditions. Again, a graphical representation can be achieved in the Nyquist plane (Fig. 5.41). The discs centred in -1 and in $G_{ol}(j\omega)$ respectively must then be disconnected to guarantee condition (5.135).

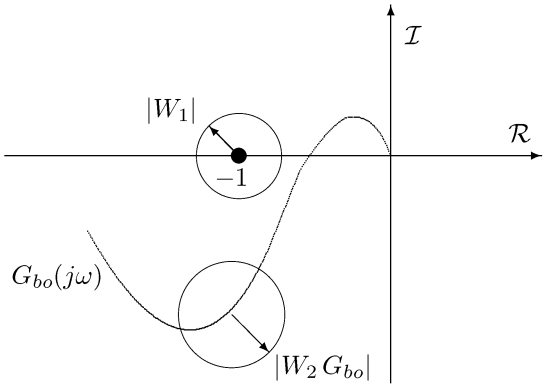


Figure 5.41: Nyquist locus for the open-loop transfer function with the performance and stability discs: these two discs must be disconnected to guarantee robust stability

These considerations for robustness find their application in the design of optimal robust controllers (Doyle et al., 1992; Oustaloup, 1994), which minimize

the following criterion deduced from Eq. (5.135)

$$\mathcal{C}(C) = \| \left(|W_1\, S|^2 + |W_2\, T|^2 \right)^{1/2} \|_\infty \tag{5.136}$$

Let us cite some general ideas and recommendations

- The spectra of signals of reference and disturbances (factor of $S(s)$ in Eq. (5.119)) are concentrated in general around low frequencies and the spectrum of measurement errors (factor of $T(s)$ in Eq. (5.119)) covers a larger frequency domain.
- In a simplified manner, the wish is to obtain small $|S(j\omega)|$ at low frequencies (in the process passband) and small $|T(j\omega)|$ at high frequencies (in the process stopband).
- The stability margin is frequently estimated by calculating the maximum of the modulus of the sensitivity function $\max(|S(j\omega)|)$ or of the complementary sensitivity function $\max(|T(j\omega)|)$. Thus, the modulus margin is introduced, for which De Larminat (1993) recommends

$$\max_\omega |T(j\omega)| \leq 1.3 \tag{5.137}$$

which is, in general, less conservative than Eq. (5.132).
- In a very similar way to Fig. 5.30, the complementary sensitivity function $T(j\omega)$ is close to 1 at low frequencies, presents a peak for a resonance frequency, and then decreases at high frequencies. T should decrease quickly at high frequencies in order to diminish the influence of model errors in this frequency domain.
- The sensitivity function $S(j\omega)$ is small at low frequencies to protect from model errors in this frequency domain, and tends towards 1 at high frequencies. If possible, S should not present a peak in the transition region.
- Control u, according to Fig. 5.38, is equal to

$$u = \frac{C}{1 + C\,P}(y_r - d - \eta) \tag{5.138}$$

The amplitude of process inputs is physically limited. When the open-loop gain $|G_{ol}| = |CP|$ is large with respect to 1, the ratio $C/(1+C\,P)$ is near $1/P$ and represents the sensitivity of the input with respect to the disturbances and to the set point, which could be denoted by S_u. In comparison with the complementary sensitivity function T, the relation $T = S_u P$ is obtained. As the input u is to be limited, the sensitivity S_u should not be too high, which is realized by imposing a large open-loop gain at low frequency and by making S_u decrease rapidly at high frequency.
- The output, following Fig. 5.38, is equal to

$$y = \frac{C\,P}{1 + C\,P}\,(y_r - \eta) + \frac{1}{1 + C\,P}\,d = T\,(y_r - \eta) + S\,\eta \tag{5.139}$$

The measurement noise η thus influences the output through the complementary sensitivity function T, hence the importance of the decrease of

$T(j\omega)$ at high frequencies. The influence of the set point is also realized through the complementary sensitivity function at low frequencies. To avoid too large inputs, it may be necessary to filter the set point, which reduces the passband.

- In the Bode plot, the open-loop transfer function must respect a certain frequency specification performance-robustness. At low frequency, the condition of nominal performance (5.123) prevails, while at high frequency, this will be the condition of robust stability (5.132).
- A necessary condition of robust performance is that

$$\min |W_1(j\omega)|, |W_2(j\omega)| < 1 \tag{5.140}$$

which implies that at least one of the two functions W_1 or W_2 has a modulus lower than 1. In practice, on the one hand, $|W_1(j\omega)|$ is a decreasing function with respect to ω dominating at low frequency for the tracking of low frequency signals, on the other, $|W_2(j\omega)|$ is an increasing function with respect to ω, dominating at high frequency as well as the uncertainty.

Design of a controller C:
The controller must satisfy the condition of robust performance (5.135). A possibility is to build the function G_{ol} so that this condition is verified, then to calculate the controller as $C = G_{ol}/P$. Of course, C must be proper. Moreover, if P or P^{-1} is not stable, G_{ol} must contain the unstable poles. With respect to G_{ol}, the condition of robust performance is written as

$$\left|\frac{W_1}{1+G_{ol}}\right| + \left|\frac{W_2}{1+G_{ol}}\right| < 1 \qquad \forall\omega \tag{5.141}$$

while respecting Eq. (5.140).

- At low frequency, $|W_1| \gg 1 > |W_2|$, which gives

$$|G_{ol}| > \frac{|W_1|}{1-|W_2|} \qquad \text{or} \qquad \frac{|W_1|}{1-|W_2|}\,|S| < 1 \tag{5.142}$$

This condition of robust performance is stronger than the condition of nominal performance $|W_1 S| < 1$.

- At high frequency, $|W_2| \gg 1 > |W_1|$, which gives

$$|G_{ol}| < \frac{1-|W_1|}{|W_2|} \qquad \text{or} \qquad \frac{|W_2|}{1-|W_1|}\,|T| < 1 \tag{5.143}$$

This condition of robust performance is stronger than the condition of robust stability $|W_2 T| < 1$.

The design of G_{ol} according to the frequency specification can then be performed in the following manner. In the Bode plot for the modulus, the curve of $(|W_1|)/(1-|W_2|)$ is drawn in the low-frequency domain and the curve of $(1-|W_1|)/(|W_2|)$ in the high-frequency domain. In fact, given the characteristics of W_1 and W_2, it is practically equivalent to simply consider the curve of $(|W_1|)$ at low frequency and the curve of $1/(|W_2|)$ at high frequency. The

curve of G_{ol} must be situated above the first curve at low frequency, and below the second one at high frequency. Moreover, so that the controller transfer function C is proper, $|G_{ol}|$ must decrease at least as fast as $|P|$ at high frequency. Lastly, the transition from low to high frequency must be smooth with a small slope when the modulus is near 1 (gain crossover frequency). Doyle et al. (1992) recommend a maximum slope equal to 2, and that at the gain crossover frequency, the argument of G_{ol} is larger than $-\pi$. A trial and error procedure will often be necessary.

Example 5.3: **Design of a Robust Controller for Delay Compensation**

Design a controller C that allows us to stabilize a first-order process presenting an unknown time delay.

Consider a first-order process, with time delay, of the transfer function

$$\tilde{P} = \frac{1}{\tau\, s + 1} \exp(-t_d s) = P \exp(-t_d s) = P\,(1 + \Delta W_2) \tag{5.144}$$

where P represents the first-order process without time delay, thus is non-perturbed. Suppose that the unknown time delay t_d is included between 0 and a maximum time delay t_{dm}. Numerical values are $\tau = 10$, $t_{dm} = 4$. The uncertainty associated with the time delay is considerable compared to the process time constant. The design of the controller will be realized by a series of trial and error calculations.

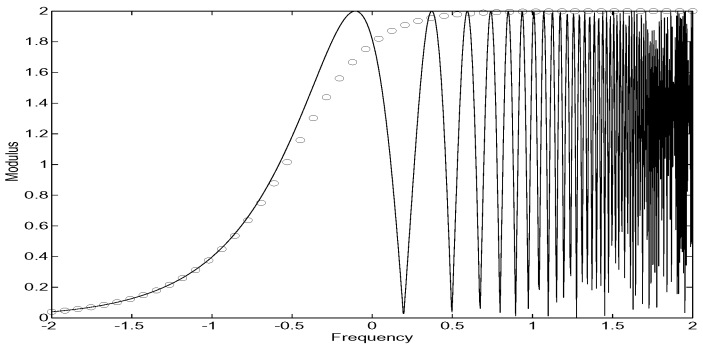


Figure 5.42: Relative uncertainty (continuous curve) and its approximation ('o') for the first-order process ($\tau = 10$) with unknown time delay ($t_{dm} = 4$)

The relative uncertainty associated with the process (Fig. 5.42) is equal to

$$\frac{\Delta W_2 P}{P} = \exp(-t_d s) - 1 \tag{5.145}$$

Replacing s by $j\omega$ gives

$$\frac{\Delta W_2 P}{P} = -2j \sin(\frac{t_d}{2}\omega) \exp(-j\frac{t_d}{2}\omega) \tag{5.146}$$

having the following modulus

$$\left|\frac{\Delta W_2 P}{P}\right| = \left|2\sin(\frac{t_d}{2}\omega)\right| \tag{5.147}$$

In fact, the time delay t_d is unknown, but is itself bounded by t_{dm} so that an approximation of the bound of the relative uncertainty is

$$\left|\frac{\Delta W_2 P}{P}\right| \leq \left|2\sin(\frac{t_{dm}}{2}\omega)\right| \tag{5.148}$$

from which the frequency weighting concerning robustness is deduced

$$\begin{cases} W_2(j\omega) = \left|2\sin(\frac{t_{dm}}{2}\omega)\right| & \text{if } \omega < \dfrac{\pi}{t_{dm}} \\ W_2(j\omega) = 2 & \text{if } \omega \geq \dfrac{\pi}{t_{dm}} \end{cases} \tag{5.149}$$

The higher limit of the bound is reached at frequency $\frac{\pi}{t_{dm}}$ when $\sin(\frac{t_{dm}}{2}\omega) = 1$. On the other hand, the uncertainty is equal to 100% when $\Delta = 1$, which gives the maximum bandwidth $[0, \frac{\pi}{3t_{dm}}] \approx [0, \frac{1}{t_{dm}}]$. Using a Padé approximation of order 1, the relative uncertainty is approximated (Fig. 5.42) by

$$\frac{\Delta W_2 P}{P} \approx \frac{t_d s}{\frac{t_d}{2}s + 1} \tag{5.150}$$

so that we choose

$$W_2 = \frac{t_{dm} s}{\frac{t_{dm}}{2}s + 1} \tag{5.151}$$

By setting $G_{ol} = \tilde{P}C$, C being the controller transfer function, the following relation will have to be checked in the high-frequency domain

$$|G_{ol}(j\omega)| \leq \frac{1}{|W_2(j\omega)|} \quad \forall \omega > \frac{1}{t_{dm}} \tag{5.152}$$

The transition for G_{ol} is located in the frequency range $[\frac{\pi}{3t_{dm}}, \frac{\pi}{t_{dm}}] \approx [\frac{1}{t_{dm}}, \frac{2}{t_{dm}}]$. Set $\omega_1 = \frac{1}{t_{dm}}$, $\omega_2 = \frac{2}{t_{dm}}$.

We begin the calculation by setting the open-loop transfer function

$$G_{ol} = \frac{b}{cs + 1} \tag{5.153}$$

b should be as large as possible to minimize the tracking error. We take $c = t_{dm}$ so that G_{ol} begins to decrease at $\frac{1}{t_{dm}}$. We also desire that $|G_{ol}| < 1$. As $|W_2|$ is equal to 1 for $\omega = 2/(\sqrt{3}t_{dm}) = 0.289$, it gives $b = 1.53$. Note that this constraint is more severe than $|G_{ol}|_{\omega_2} < 1$.

On the other hand, the frequency weighting relative to performance is chosen as

$$\begin{cases} |W_1| = a & \text{if } \omega < \frac{1}{t_{dm}} \\ |W_1| \approx 0 & \text{if } \omega \geq \frac{1}{t_{dm}} \end{cases} \tag{5.154}$$

From the inequality valid at low frequencies

$$|G_{ol}| \geq \frac{|W_1|}{1-|W_2|} \quad \forall \omega < \omega_1 \tag{5.155}$$

it results that

$$|G_{ol}(j\omega_1)| = \frac{a}{1-|W_2(j\omega_1)|} \tag{5.156}$$

which gives $a = 0.316$.

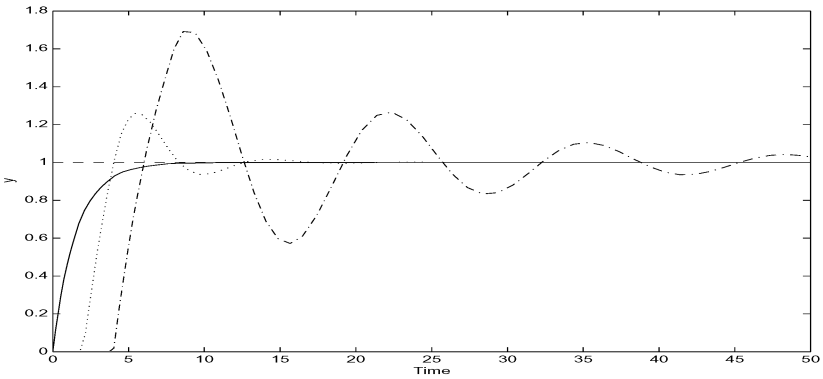


Figure 5.43: Response to a set point step of the perturbed system for different values of the time delay (0, 2 and 4) and the first controller

For the first-order process chosen, fixing the time delay (0, 2 and 4), it is then possible to simulate the closed-loop response to a set point step (Fig. 5.43). As the controller equal to

$$C = \frac{1.53\,(10s+1)}{4s+1} \tag{5.157}$$

presents no integral action, the deviation which existed with respect to the set point has been compensated for by the introduction of a pure gain equal to 1.654 between the set point and the comparator. For the largest value of the time delay, the response presents a relatively large overshoot. Thus, this constitutes the first trial with the following numerical value of the open-loop transfer function G_{ol}

$$G_{ol} = \frac{1.53}{4s+1} \tag{5.158}$$

We wish to increase $|G_{ol}|$ on the passband $[0, \frac{1}{t_{dm}}]$ by modifying G_{ol} as

$$G_{ol} = \frac{b}{4s+1}\,\frac{s+5}{s+1} \tag{5.159}$$

The same method as previously gives $b = 0.42$. The weighting W_2 is unchanged, as well as $W_1 = a$. If the value $b = 0.42$ is kept, the response is more unstable

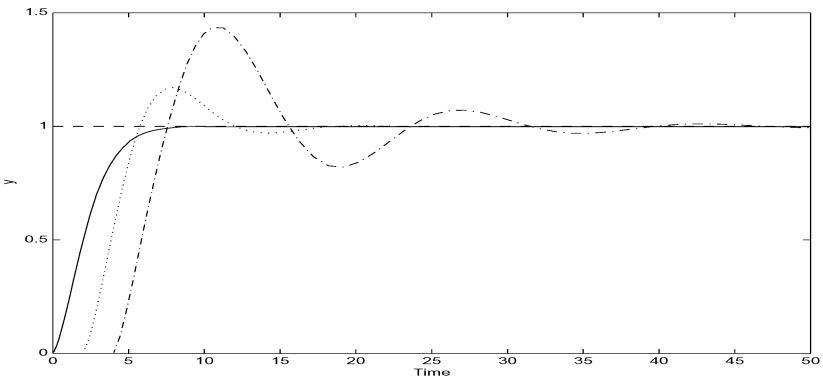


Figure 5.44: Response to a set point step of the perturbed system for different values of the time delay (0, 2 and 4) and the second controller

than in the previous case. For that reason, b has been modified arbitrarily to $b = 0.21$, which allows the correct responses as in Fig. 5.44. The precompensation gain is equal to 1.95.

This constitutes the second trial, with the following controller transfer function

$$C = \frac{0,21\,(10s+1)}{4s+1}\,\frac{s+5}{s+1} \tag{5.160}$$

The waited improvement with respect to the previous controller, which resulted from the simple application of robustness rules, is not evident.

For the third calculated controller, the choice is made to provide it with an integral action so that its transfer function is

$$C = \frac{b(10s+1)}{s}\,\frac{s+5}{s+1} \tag{5.161}$$

and the open-loop transfer function is

$$G_{ol} = \frac{b}{s}\,\frac{s+5}{s+1} \tag{5.162}$$

The calculation then gives $b = 0.08$ and $W_1 = a = 0.338$. With this value, oscillations are strong, so b is reduced to $b = 0.04$. With this value, responses are quite acceptable (Fig. 5.45), and it is not necessary anymore to set a precompensation gain. This constitutes the third trial.

By filtering the set point by a first-order transfer function equal to $1/(10s+1)$, it is possible to absorb much larger time delays than t_{dm}. Thus, the response of Fig. 5.46 was obtained for $t_d = 6$ at the expense of a sluggish oscillation but with a smaller overshoot than the response obtained with $t_d = 4$ and precompensations by only pure gains.

Figure 5.47 shows that in these conditions, the modulus $|G_{ol}|$ of the open-loop transfer function is larger than $|W_1|$ at low frequencies respecting the

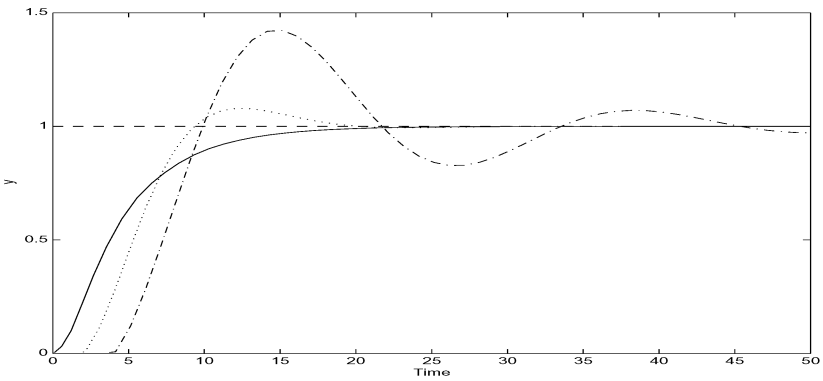


Figure 5.45: Response to a set point step of the perturbed system for different values of the time delay (0, 2 and 4) and the third controller with integral action

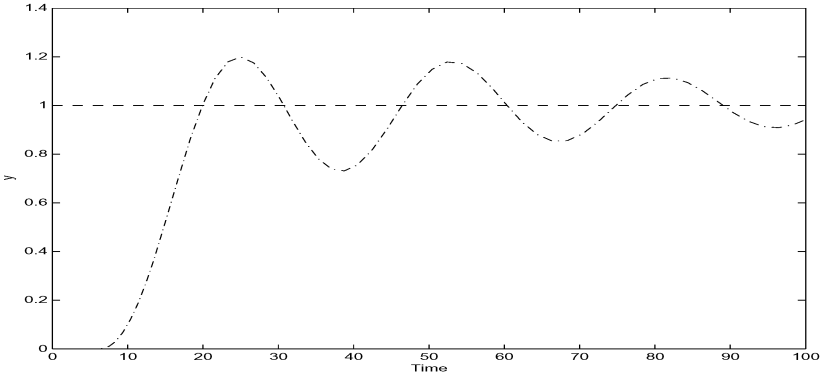


Figure 5.46: Response to a set point step of the perturbed system for a time delay equal to 8 and a filtering of the set point by a first-order

nominal performance, and is smaller than $1/|W_2|$ at high frequencies respecting the robust stability. The set $|W_1|$ and $1/|W_2|$ gives the frequency specification to be respected. Moreover, the sum $|W_1 S| + |W_2 T|$ relative to the condition of robust performance is lower than 1 on all the frequency range. The robustness criteria are thus respected.

If the behaviour of the sensitivity function S and of the complementary sensitivity function T (Fig. 5.48) is examined, it shows that the modulus of T tends towards 1 at low frequencies. At high frequencies, the modulus of S tends towards 1 as foreseen. In the intermediate frequency range, the modulus of S shows a small bump while $|T|$ decreases continuously. It would be better that $|T|$ presents a slight bump at the resonance frequency.

Doyle et al. (1992) frequently recommend using

$$|W_1| = 0.5\,(|P|^2 + 1)^{1/2}\,, \quad |W_2| = 0.5\,(|P|^{-2} + 1)^{1/2} \tag{5.163}$$

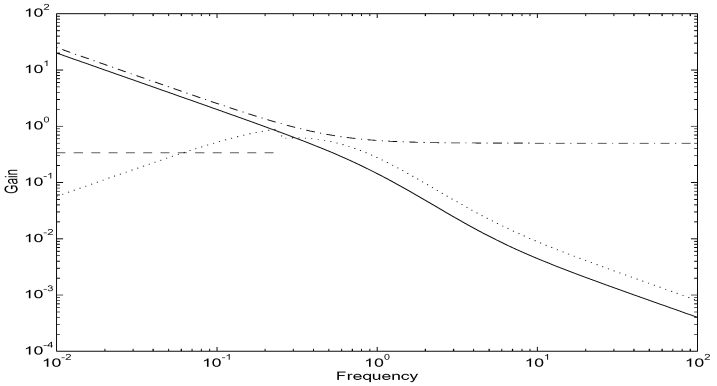


Figure 5.47: Modulus $|W_1|$ of the weighting function relative to nominal performance (– –). Reciprocal $1/|W_2|$ of the modulus of the weighting function relative to robust stability (– - –). Modulus $|G_{ol}|$ of the open-loop transfer function (continuous curve). Sum of moduli $|W_1S| + |W_2T|$ for robust performance (. . .)

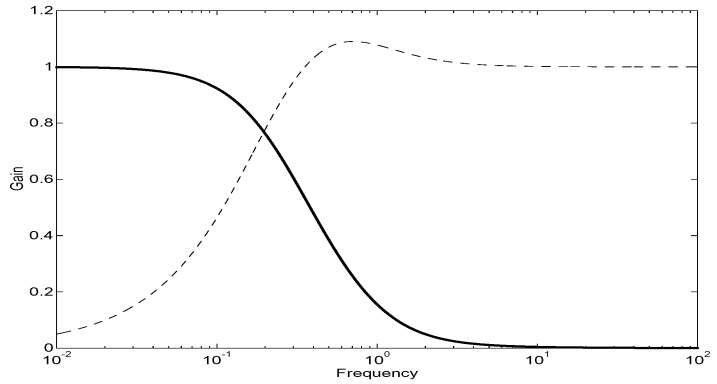


Figure 5.48: Modulus of sensitivity function S (- -) and of complementary sensitivity function T (continuous curve)

In the present case, it gives

$$P = \frac{K}{\tau s + 1} \Longrightarrow \begin{cases} |W_1| \approx 0.5(K^2+1)^{1/2} , & \forall \omega < \omega_1 \\ |W_2| \approx 0.5 \dfrac{\tau}{K} \omega , & \forall \omega > \omega_2 \end{cases} \tag{5.164}$$

The new weightings have been drawn in Fig. 5.49 which must be compared to Fig. 5.47. It shows that the modulus $|W_1|$ is increased and the allure of $1/|W_2|$ is relatively different and imposes a more severe constraint. The

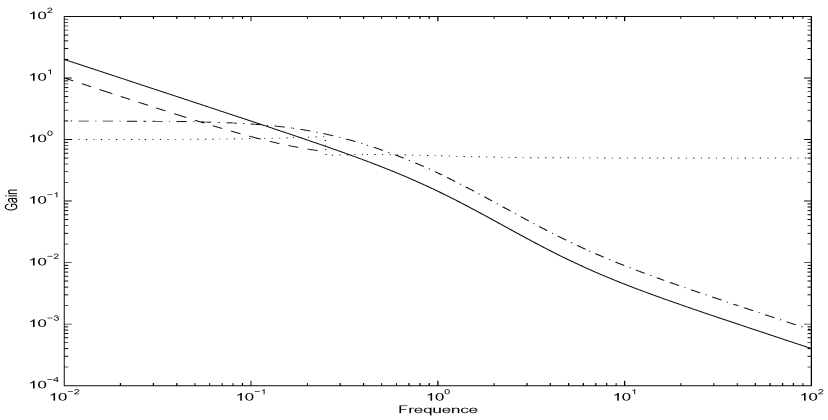


Figure 5.49: Application of recommendations of Doyle et al. (1992). Modulus $|W_1|$ of the weighting function relative to nominal performance (– –). Reciprocal $1/|W_2|$ of the modulus of the weighting function relative to robust stability (– - –). Modulus $|G_{ol}|$ of the open-loop transfer function (continuous curve). Sum of moduli $|W_1 S| + |W_2 T|$ for robust performance (. . .)

frequency specification is still respected by the same open-loop transfer function as previously, thus the same controller.

Kwakernaak in (Oustaloup, 1994), setting $L = G_{ol}$ and $L^{-1} = G_{ol}^{-1}$, recommends choosing the following weightings

$$\begin{cases} |W_1(j\omega)| = \begin{cases} |W_{L^{-1}}(j\omega)| & \text{at low frequencies} \\ 0 & \text{at high frequencies} \end{cases} \\ |W_2(j\omega)| = \begin{cases} 0 & \text{at low frequencies} \\ |W_L(j\omega)| & \text{at high frequencies} \end{cases} \end{cases} \tag{5.165}$$

where functions $|W_L|$ and $|W_{L^{-1}}|$ respectively represent bounds of relative uncertainties on the open-loop transfer function and its reciprocal.

Taking the controller $C = 1$, one obtains $G_{ol}^{-1} = (1 + \tau s)\exp(t_d s)$, which gives at low frequencies

$$W_{L^{-1}}(s) = 1 - \exp(-t_d s) \quad \Longrightarrow \quad W_1(j\omega) \approx j t_d \omega \tag{5.166}$$

while, at high frequencies, the same expression as at the beginning of this example is obtained

$$W_L(s) = \exp(-t_d s) - 1 \quad \Longrightarrow \quad W_2(j\omega) = \exp(-j t_d \omega) - 1 \tag{5.167}$$

The weighting W_1 reaches its maximum at low frequencies (domain $[0, 1/t_m]$) in $\omega_1 = 1/t_{dm} = 0.25$, that gives $|W_1| = 1$, which is not a very large value, neither is it very different from previously found values.

5.11 Summary for Controller Design

Maciejowski (1989) gives an excellent description of a controller design. Refer to Fig. 5.38 for notations.

In a simplified way, the desire is that the sensitivity function $|S(j\omega)|$, which intervenes as a factor of disturbances or of parameter uncertainty is small at low frequencies ($\omega < \omega_o$) and the complementary sensitivity function $|T(j\omega)|$, which intervenes as a factor of the measurement noise is small at high frequencies ($\omega > \omega_b$). The design results from a compromise.

The classical rule is that to make $|S(s)|$ small, $|1 + C(s)P(s)|$ should be large, thus if $|P(s)|$ is not large, a large controller gain $|C(s)|$ is necessary at low frequencies. Similarly, to make $|T(s)|$ small, $|1 + C(s)P(s)|$ should be near zero; it may be necessary that the controller gain $|C(s)|$ is small at high frequencies, when $|P(s)|$ does not decrease sufficiently fast at high frequencies.

The frequency ω_o, is defined taking into account the spectrum of the disturbance, as

$$|S(j\omega)| < \epsilon \qquad , \qquad \omega \leq \omega_o \tag{5.168}$$

The frequency ω_b is the passband of the loop and is the smallest frequency such that

$$|T(j\omega_b)| = T(0)/\sqrt{2} \tag{5.169}$$

It is larger than ω_o. The passband ω_b is approximately reciprocal with respect to the response time of the output $y(t)$ to a disturbance step. Thus, it is an indicator of the speed of rejection of a disturbance. On the other hand, the passband ω_b is frequently very near to the gain crossover frequency ω_g and is often such that

$$\omega_\phi \leq \omega_b \leq 2\,\omega_g \tag{5.170}$$

The transmission band equal to the passband of $F(s)T(s)$ indicates the speed of the response to the set point.

The internal model principle by Francis and Wonham (1976) must be used. It can be expressed in frequency terms such that

$$|P(0)C(0)| = \infty \tag{5.171}$$

to ensure a zero steady-state error facing step disturbances, or

$$\lim_{\omega \to 0} \omega |P(j\omega)C(j\omega)| = \infty \tag{5.172}$$

to ensure a zero steady-state error facing ramp disturbances.

The loop stability must be considered, an excessive phase lag must be avoided when

$$|P(j\omega)C(j\omega)| = 1 \tag{5.173}$$

because of the Nyquist criterion, which limits the action on the loop gain. Near the gain crossover frequency, the speed of decrease of the open-loop transfer function, for a minimum-phase system, must be lower than 40dB/decade.

It is better to design a stable open-loop controller, although this is not essential.

The existence of unstable zeros or poles in process $P(s)$ diminishes the efficiency of the output feedback by decreasing the frequency domain on which it is efficient.

For a stable open-loop system, the Bode criterion gives

$$\int_0^\infty \log(|S(j\omega)|d\omega = 0 \tag{5.174}$$

In fact, if the sensitivity decreases more strongly in a domain of low frequencies, a sensitivity peak results at higher frequencies.

References

P. De Larminat. *Automatique, Commande des Systèmes Linéaires.* Hermès, Paris, 1993.

J.C. Doyle, B.A. Francis, and A.R. Tannenbaum. *Feedback Control Theory.* Maxwell Macmillan, New York, 1992.

B.A. Francis and W.M. Wonham. The internal model principle of control theory. *Automatica*, 12:457–465, 1976.

H. Kwakernaak. Robust control and $\mathcal{H}\infty$-optimization – tutorial paper. *Automatica*, 29(2):255–273, 1993.

P. De Larminat. La commande robuste: un tour d'horizon. *APII*, 25:267–296, 1991.

J.M. Maciejowski. *Multivariable Feedback Design.* Addison-Wesley, Wokingham, England, 1989.

M. Morari and E. Zafiriou. *Robust Process Control.* Prentice Hall, 1989.

A. Oustaloup, editor. *La Robustesse. Analyse et Synthèse de Commandes Robustes.* Hermès, Paris, 1994.

S.M. Shinners. *Modern Control System Theory and Design.* Wiley, New York, 1992.

W.M. Wohnam. *Linear Multivariable Control. A Geometric Approach.* Springer-Verlag, New York, 1985.

Chapter 6

Improvement of Control Systems

Three advanced control systems have been presented in Chap. 4: internal model control, pole-placement and linear quadratic control. These control systems are essentially implementable in a computer or a microprocessor. Excepted these cases, the only type of control studied up to this chapter concerned simple feedback closed-loop with proportional, PI or PID controllers. Although this type of control is very important and very much used, it may be insufficient for all existing systems because of different problems met in real processes.

6.1 Compensation of Time Delay

Time delays inside a process control system may have various origins: transport of fluids, phenomena that are slow to appear, recycle loops and dead times due to measurement devices. In all of these cases, classical control will not be satisfactory. Thus, a disturbance is detected only a long time after it effectively enters the system: indeed, it is necessary that the measured output has deviated with respect to the set point; therefore, the action to compensate the disturbance effect will be performed with a time delay and often will be inappropriate.

The pure time delay, by adding a phase lag to the feedback loop, is an important source of instability for closed-loop control. The tendency will then be to reduce the controller gain, which will lead to a more sluggish response.

A solution to this problem is the Smith predictor ; the development of this predictor lies in the knowledge of a process model (Smith, 1957, 1959). The model of the process transfer function $\tilde{G}_p(s)$ is decomposed into two factors

$$\tilde{G}_p(s) = \tilde{G}(s)\ \exp(-\tilde{t_d}s) \tag{6.1}$$

i.e. a classical transfer function and a pure time delay.

The theoretical open-loop response (Fig. 6.1) to a set point variation is equal to

$$Y_m(s) = K_m\, G_c(s)\, \tilde{G}_a(s)\, \tilde{G}(s)\, \exp(-\tilde{t}_d s)\, \tilde{G}_m(s)\, Y_r(s) \tag{6.2}$$

where the "tilde" notation indicates a modelled variable or function (thus including an error). The real response is simply deduced from the theoretical response by suppressing the "tilde".

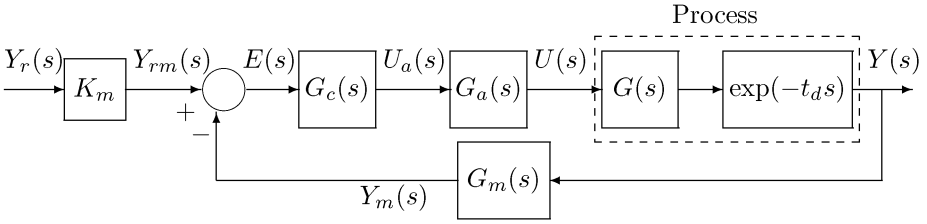


Figure 6.1: Feedback closed-loop system in the case of a process with time delay

Thus, the measured variable $Y_m(s)$ is known to the user with a time delay.

The aim of the Smith predictor is, based on the knowledge of the time delay, to operate as if the process transfer function was effectively separated into two factors and thus to obtain the information $Y^*(s)$ without time delay, as demonstrated in Fig. 6.2 which is completely theoretical. As a matter of fact, it is impossible to physically separate the time delay from the rest of the process.

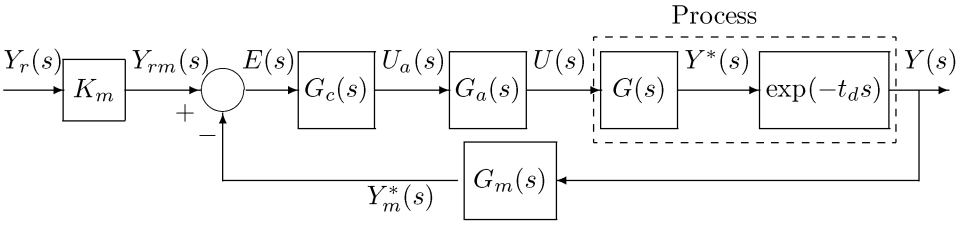


Figure 6.2: Theoretical obtaining of the measured output without time delay

The theoretical measured output without time delay would be equal to

$$Y_m^*(s) = K_m\, G_c(s)\, \tilde{G}_a(s)\, \tilde{G}(s)\, \tilde{G}_m(s)\, Y_r(s) \tag{6.3}$$

this represents the desired output; it then "suffices" to add to $Y_m(s)$ the quantity

$$K_m\, G_c(s)\, \tilde{G}_a(s)\, \tilde{G}(s)\, [1 - \exp(-\tilde{t}_d s)]\, \tilde{G}_m(s)\, Y_r(s) \tag{6.4}$$

to get that result. Imagine that, in the block diagram of Fig. 6.3, the wire bearing the signal y_m^* is cut (the loop is opened); in the real process (in open

loop), the following signal is obtained

$$\begin{aligned} Y_m^{*'} &= \tilde{G}_a\,\tilde{G}\,\tilde{G}_m\,(1-\exp(-\tilde{t}_d s))\,G_c\,K_m\,Y_r + K_m\,G_c\,G_a\,G\,\exp(-t_d s)\,G_m\,Y_r \\ &\approx G_a\,G\,G_m\,G_c\,K_m\,Y_r \end{aligned} \tag{6.5}$$

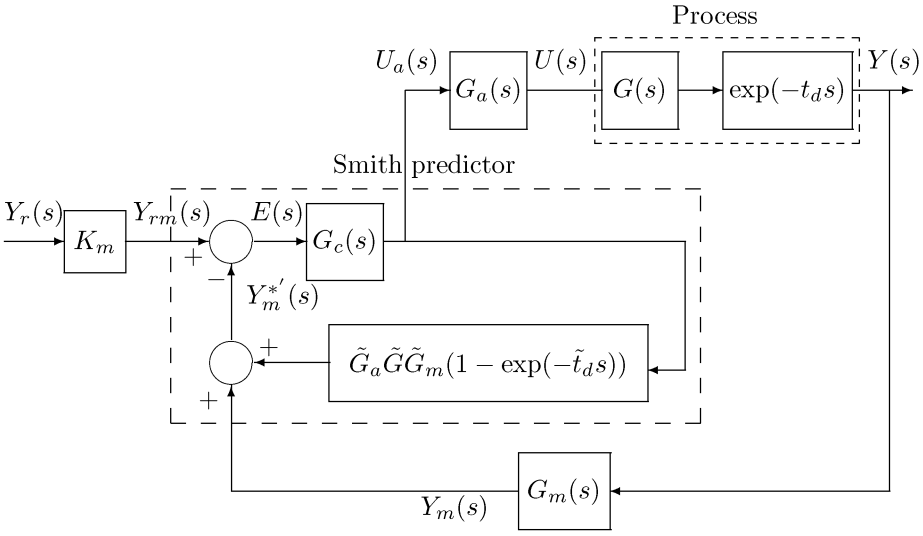


Figure 6.3: Smith predictor

The Smith predictor is thus composed of the conventional controller and a dead time compensator, which plays the prediction role. Internal model control can be mentioned concerning this controller, as the process model parameters are used in the predictor.

The compensation will be perfect only if the model itself is perfect and the most important deals with the time delay. Thus, Schleck and Danesian (1978) found that when the error is larger than 30% the response is no better than with a classical controller. Moreover, disturbances pose serious problems (VanDoren, 1996).

The realization of an analog Smith predictor is not simple and many modified Smith controllers have been proposed to improve the original one. On the other hand, its discrete equivalent is easy to realize and, moreover, the latter can be made adaptive (Niculescu and Annaswamy, 2003).

6.2 Inverse Response Compensation

The phenomenon of inverse response is produced when two antagonistic effects occur at the core of a process. The response initially goes in one direction as a transfer function overcomes the other one for small periods of time, then goes in the opposite way towards its steady state when the second transfer function

overcomes the first one. Such a system (Fig. 6.4) has already been presented previously for two first-order transfer functions.

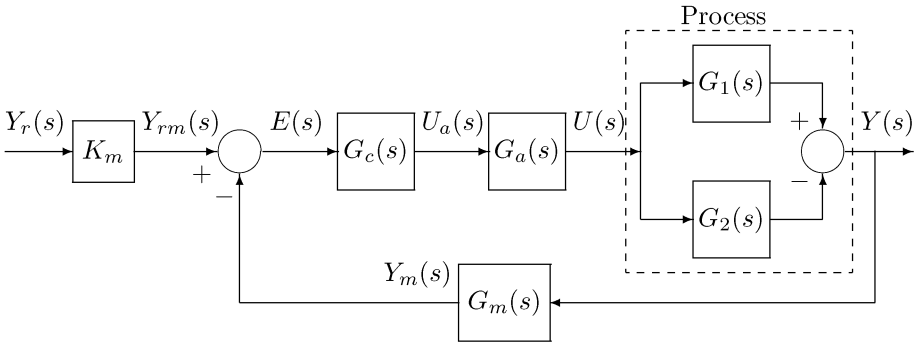


Figure 6.4: Inverse response process

It is possible with a PID controller to limit the influence of the inverse response owing to the feedforward effect of the derivative action.

The best compensation that could be simply realized is based on a compensator designed in the same spirit as the Smith predictor.

The following reasoning is performed for two first-order systems and valid only in that case, but the general principle remains identical.

Let

$$G_1(s) = \frac{K_1}{\tau_1 s + 1} \quad \text{and} \quad G_2(s) = \frac{K_2}{\tau_2 s + 1} \tag{6.6}$$

be the two considered first-order transfer functions representing the process (Fig. 6.5). The condition of inverse response (existence of a positive zero) is

$$\frac{\tau_1}{\tau_2} > \frac{K_1}{K_2} > 1 \tag{6.7}$$

The Laplace transform of the open-loop output is equal to

$$\begin{aligned} Y &= G_c\, G_a\, [G_1 - G_2]\, K_m\, Y_r \\ &= G_c\, G_a \left[\frac{K_1(\tau_2 s + 1) - K_2(\tau_1 s + 1)}{(\tau_1 s + 1)\,(\tau_2 s + 1)} \right] K_m\, Y_r \\ &= G_c\, G_a \left[\frac{(K_1\,\tau_2 - K_2\,\tau_1)s + (K_1 - K_2)}{(\tau_1 s + 1)\,(\tau_2 s + 1)} \right] K_m\, Y_r \end{aligned} \tag{6.8}$$

The transfer function Y/Y_r possesses a positive zero (thus, the inverse response will occur) when

$$s = \frac{K_2 - K_1}{K_1\,\tau_2 - K_2\,\tau_1} > 0 \tag{6.9}$$

To eliminate the inverse response, it "suffices" to suppress this positive zero of the open-loop response. The quantity $Y'(s)$ is added to $Y(s)$ so that the desired output

$$Y^*(s) = Y'(s) + Y(s) \tag{6.10}$$

does not present a positive zero anymore. It gives

$$Y'(s) = a\,G_c\,G_a\,K_m\left[\frac{G_2}{K_2} - \frac{G_1}{K_1}\right]Y_r \tag{6.11}$$

and

$$\begin{aligned} Y^*(s) &= G_c\,G_a\,K_m\left[G_1 - G_2 + \frac{a\,G_2}{K_2} - \frac{a\,G_1}{K_1}\right]Y_r \\ &= G_c\,G_a\,K_m\left[\frac{K_1 - a}{\tau_1 s + 1} - \frac{K_2 - a}{\tau_2 s + 1}\right]Y_r \\ &= G_c\,G_a\,K_m\left\{\frac{[(K_1 - a)\,\tau_2 - (K_2 - a)\,\tau_1]\,s + (K_1 - K_2)}{(\tau_1 s + 1)\,(\tau_2 s + 1)}\right\}Y_r \end{aligned} \tag{6.12}$$

The necessary and sufficient condition for a zero to be negative is thus

$$s = -\frac{K_1 - K_2}{(K_1 - a)\,\tau_2 - (K_2 - a)\,\tau_1} < 0 \tag{6.13}$$

As $K_1 > K_2$, the condition on parameter a results

$$a > \frac{K_2\,\tau_1 - K_1\,\tau_2}{\tau_1 - \tau_2} \tag{6.14}$$

The compensator transfer function is thus equal to

$$G_{\text{comp}} = a\,\tilde{G}_a\left[\frac{1}{\tilde{\tau}_2 s + 1} - \frac{1}{\tilde{\tau}_1 s + 1}\right]\tilde{G}_m \tag{6.15}$$

taking into account the models of the actuator denoted by a "tilde", of the process and of the measurement.

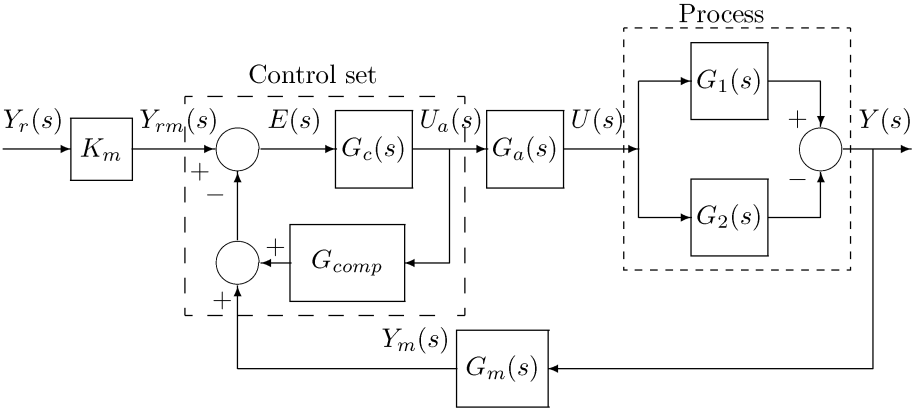


Figure 6.5: Compensation of inverse response

The block diagram corresponding to this compensator is given by Fig. 6.5. In this case as for the Smith predictor, the efficiency of the compensator will be linked to the accuracy of the process model.

6.3 Cascade Control

Control systems studied up to now have considered only one measured output and one control variable with a single loop. In cascade control, one control variable and more than one measurement are used. A frequent example of use of cascade control is the temperature control of a chemical reactor equipped with a jacket, e.g. for exothermic reactions (Fig. 6.6). Among the possible disturbances are the feed flow rate and temperature, the thermal fluid temperature and, of course, the reaction enthalpy and kinetics.

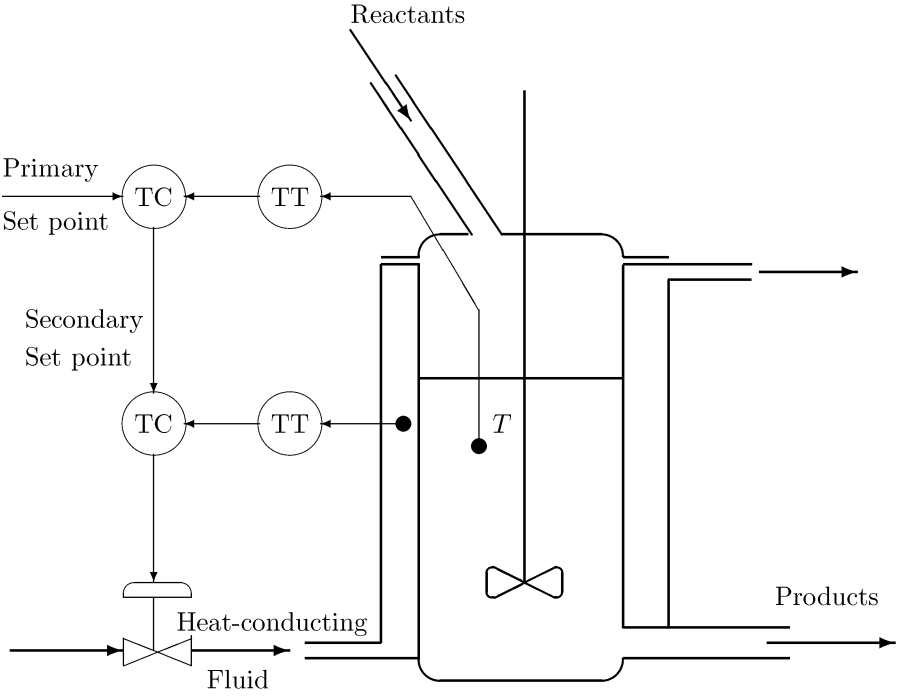


Figure 6.6: Thermal control of a reactor with jacket by cascade control

In a simple feedback control, to regulate the temperature in the reactor where an exothermic reaction occurs (inducing a potential runaway of the reaction, for example), the temperature would be measured inside the reactor; this temperature would be compared to a set point value and the inlet temperature or the flow rate of the heating-cooling fluid would be manipulated to maintain the reactor temperature close to its set point by acting through a controller on a valve situated on the thermal fluid.

When the action bears on the fluid flow rate or its inlet temperature, this action has no immediate consequence because of the heat transfer dynamics through the reactor wall and because of the thermal inertia of the reactor. To avoid this drawback, the control system can be greatly improved by realizing two temperature measurements: one in the reactor and the other one in the jacket (Fig. 6.6). The measurement in the jacket will react much faster than

the one in the reactor.

The main measurement is the temperature in the reactor; this loop is called "master" or primary. The temperature measured in the reactor is transmitted by a temperature transmitter (TT) to a controller (TC), which receives an external set point (provided by the operator). The cooling fluid disturbances are detected earlier owing to the measurement in the jacket, which acts through a second temperature transmitter (TT) on a second controller (TC); the controller of this loop has as a set point the output of the primary controller. For this reason, this loop which uses the jacket temperature is called "slave" or secondary.

Thus, two temperature measurements are available, but only one manipulated variable: the thermal fluid flow rate.

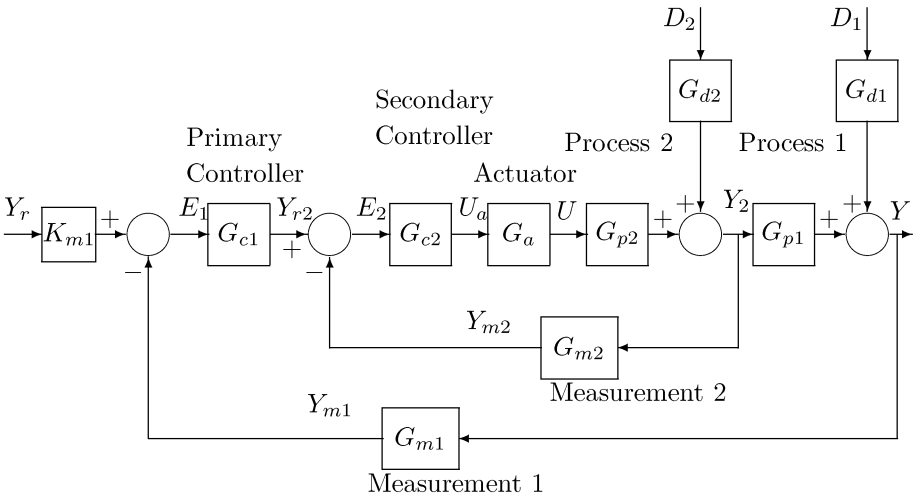


Figure 6.7: Block diagram in cascade control

According to the block diagram (Fig. 6.7), the process is divided into two parts: one affected by secondary control and the other one by primary control. The output y of process 1 is the variable to control. The output of process 2 influences process 1, but its control is not the fundamental point. Disturbances d_2 occurring at the level of the secondary loop are corrected before reaching the primary loop; this is not the case for disturbances d_1 which directly influence the primary loop.

A pure gain K_{m1} has been introduced to compensate the measurement of the primary loop. It is not necessary to use a gain K_{m2} to compensate the measurement of the secondary loop, as it can be integrated in the secondary controller gain.

The closed-loop transfer function for the secondary loop is

$$\frac{U_1}{Y_{r2}} = \frac{G_{c2}\, G_a\, G_{p2}}{1 + G_{c2}\, G_a\, G_{p2}\, G_{m2}} \tag{6.16}$$

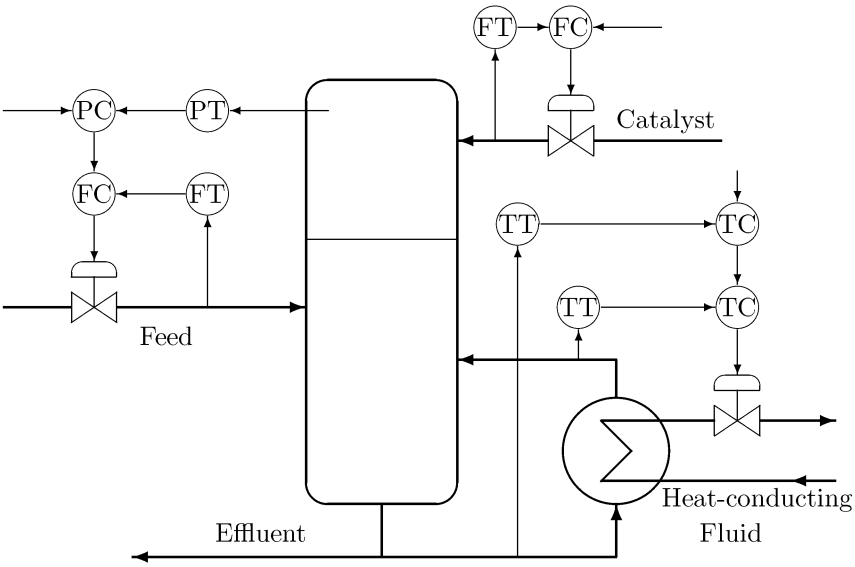


Figure 6.8: Catalytic reactor presenting two cascade controls

and the stability of the secondary loop is determined by the roots of the following characteristic equation

$$1 + G_{ol2} = 1 + G_{c2}\, G_a\, G_{p2}\, G_{m2} = 0 \tag{6.17}$$

thus the open-loop transfer function of the secondary loop is equal to

$$G_{ol2} = G_{c2}\, G_a\, G_{p2}\, G_{m2} \tag{6.18}$$

In fact, the output Y_2 of the secondary loop is the input U_1 of the primary process. The set point Y_{r2} of the secondary loop is the output of the controller of the primary loop. Finally, the cascade system can be represented as the equivalent block diagram in Fig. 6.9.

The closed-loop transfer function for the primary loop is

$$\frac{Y}{Y_r} = \frac{K_{m1}\, G_{c1}\, \dfrac{G_{c2}\, G_a\, G_{p2}}{1 + G_{c2}\, G_a\, G_{p2}\, G_{m2}}\, G_{p1}}{1 + G_{c1}\, \dfrac{G_{c2}\, G_a\, G_{p2}}{1 + G_{c2}\, G_a\, G_{p2}\, G_{m2}}\, G_{p1}\, G_{m1}} \tag{6.19}$$

and the stability of the primary loop is determined by the roots of the following characteristic equation

$$1 + G_{ol1} = 1 + G_{c1}\, \frac{G_{c2}\, G_a\, G_{p2}}{1 + G_{c2}\, G_a\, G_{p2}\, G_{m2}}\, G_{p1}\, G_{m1} = 0 \tag{6.20}$$

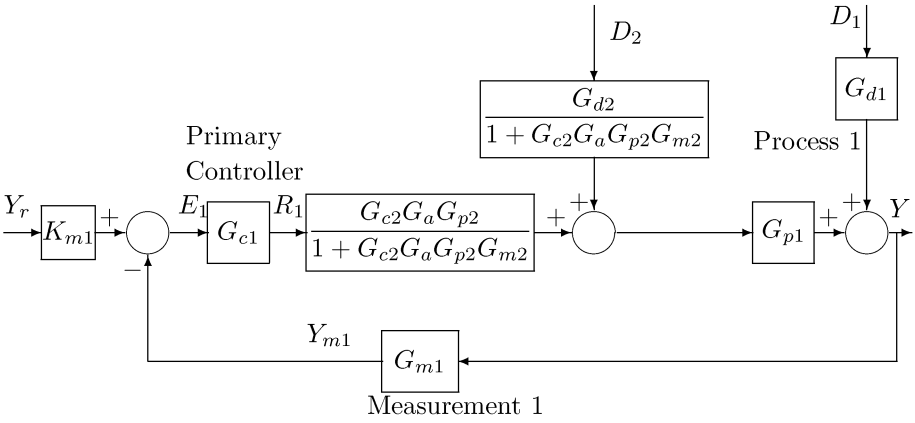


Figure 6.9: Equivalent block diagram for cascade control

thus the open-loop transfer function of the primary loop is equal to

$$G_{ol1} = G_{c1} \frac{G_{c2}\, G_a\, G_{p2}}{1 + G_{c2}\, G_a\, G_{p2}\, G_{m2}} G_{p1}\, G_{m1} \tag{6.21}$$

In general, cascade control presents better stability characteristics than a simple feedback control; moreover, it is less sensitive to modelling errors.

The secondary process presents a smaller time constant than the primary process, so that frequently the dynamics of the secondary process is neglected and it is simply represented by a pure gain (either G_{p2} or G_{d2}). The only remaining dynamics in the process is then present in G_{p1} and G_{d1}.

Cascade control is realized by two conventional controllers, but the control of the secondary loop does not need to be realized as well as that of the primary loop. For the secondary loop, a proportional controller can be used (which will produce a steady-state deviation compensated for in the primary loop), or a PI controller which will suppress this deviation. The secondary loop reacts much faster than the primary loop. Thus its phase lag is smaller and the critical frequency ω_c, if it exists, is larger. Larger gains can thus be used.

The controllers tuning is done according to the following procedure:

- First, tune the secondary controller by a classical method (e.g. Cohen-Coon or Ziegler-Nichols); make sure whether a P or PI controller is necessary.
- From the Bode plot of the global system, determine the critical crossover frequency ω_c by using the previous tuning of the secondary controller. Then, from the frequency response techniques, tune the primary controller.

In nearly all chemical processes, flow rate control is realized by cascade (e.g. heat exchanger, distillation column, absorption column)

Cascade control is frequently used in industrial processes. Figure 6.8 represents a liquid phase catalytic reactor. In one step, the pressure above the liquid phase is controlled by a cascade control by measuring as a secondary variable the feed flow rate. In the other step, the temperature in the reactor

is controlled by a second cascade control by measuring as a secondary variable the fluid temperature exiting the heat exchanger.

Example 6.1: **Cascade Control**

The simulated process is decomposed into two dynamic parts presenting time

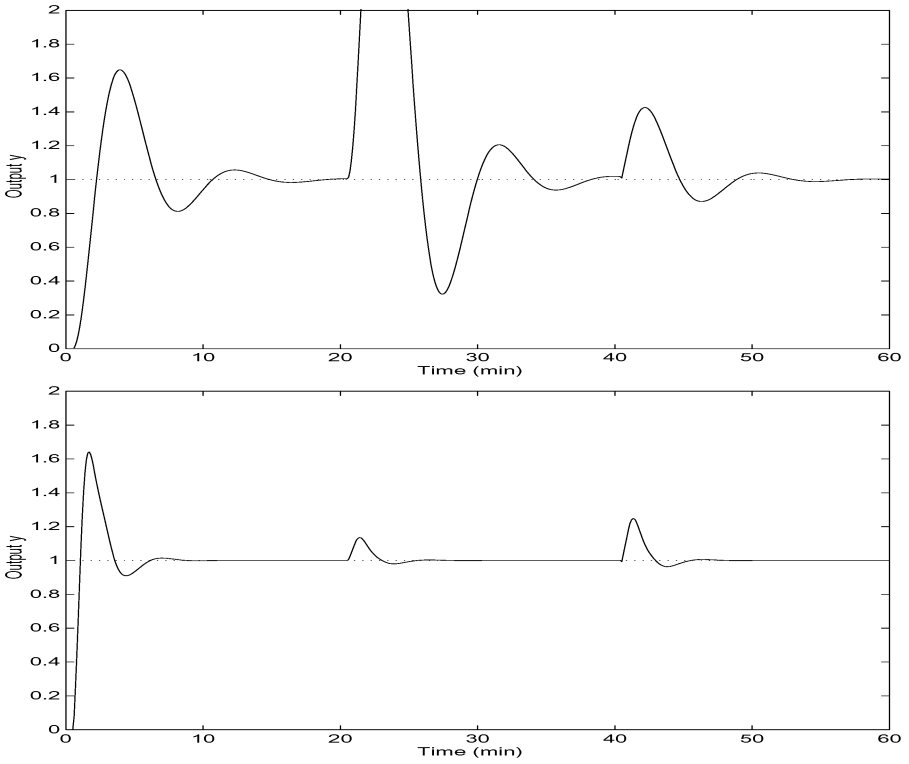


Figure 6.10: Process response in a control without cascade (top figure) or with cascade (bottom figure) to a unit step set point (at $t = 0$), a step disturbance of amplitude 0.1 on the secondary loop (at $t = 20$) and a unit step disturbance on the primary loop (at $t = 40$)

constants of different orders of magnitude, small for the secondary process and larger for the primary process. Moreover, the primary process presents a time delay. Two disturbances affect the process: d_1 inside the primary loop and d_2 inside the secondary loop. The influence of the cascade is studied by comparing the response with and without the cascade. The tuning of the primary controller is realized in both cases according to Ziegler-Nichols, Table 4.2. The data are as follows

$$G_{p2} = \frac{10}{s+1}\ ,\ G_{d2} = \frac{10}{s+1}\ ,\ G_{m2} = 0.25\ ,\ G_a = 5$$

$$G_{p1} = \frac{30\ \exp(-0.5\,s)}{10\,s+1}\ ,\ G_{d1} = \frac{4\ \exp(-0.5\,s)}{10\,s+1}\ ,\ G_{m1} = 0.1$$

where the time unit is minutes.

In the case without the cascade, Ziegler-Nichols tuning provides a PID controller that has the following characteristics

$$K_c = 0.072\ ,\ \tau_I = 2.25\text{min}\ ,\ \tau_D = 0.56\text{min} \tag{6.22}$$

In the case with the cascade, a proportional controller is chosen for the secondary loop. The gain can be freely chosen, as the secondary loop presents no crossover frequency. Nevertheless, as the steady-state gain of the open-loop transfer function of the secondary loop is equal to 12.5, a controller proportional gain equal to 1 is considered sufficient. For the primary loop, a PID controller is chosen and tuned by a Ziegler-Nichols, which gives the following characteristics

$$K_c = 1.546\ ,\ \tau_I = 1.12\text{min}\ ,\ \tau_D = 0.28\text{min} \tag{6.23}$$

This system is successively subjected to:

- At $t = 0$, a unit step set point,
- At $t = 20$ min, a step disturbance d_1 of amplitude 0.1,
- At $t = 40$ min, a unit step disturbance d_2.

Figure 6.10 shows that the presence of the cascade is visible as expected, as the influence of disturbance d_2 affecting the secondary loop is strongly decreased, the influence of disturbance d_1 is also reduced, the dynamics of the response to a set point variation is faster and the overshoots are less pronounced. The cascade dampens the disturbances and provides a better response to a set point variation.

6.4 Selective Control

In a few cases, the process presents more controlled variables than manipulated variables. In principle, the number of manipulated variables should be larger or equal to the number of outputs. It becomes clear that it will be necessary to select the outputs that will realize the control. This is performed by a set of hierarchical rules imposed by the user.

Example 6.2: **Protection of a Boiler**

Consider a boiler heated by a coil fed by vapour. If the coil is dry, even partially, resulting hot spots can lead to coil destruction. In the case of an electrical resistance, the problem is identical. Thus, there are two actions that need to be realized: the control of produced vapour pressure and the safety action, i.e. the level control in the boiler. A special system allowing passage from one measurement to another will be necessary.

Example 6.3: **Protection of a Tubular Catalytic Reactor**

The protection of a tubular catalytic reactor against the appearance of hot spots that can cause deterioration of the catalyst qualities necessitates measurement of the temperature at numerous places along the catalyst bed. On the basis of these measurements, a selector can choose the highest value and base the control on this. Moreover, the redundancy of measurements allows greater confidence in the obtained outputs. This problem is common for fluidized catalytic crackers (FCC) in refineries.

Example 6.4: **Control of a Distillation Column Reboiler**

The control variable is the heat content brought to the reboiler. There exists a lower limit on this heat quantity: below this limit, there is not enough vapour produced and the liquid does not remain on the trays; there exists a higher limit: above this limit, flooding occurs in the column.

6.5 Split-Range Control

In this case, there exists only one measurement and several manipulated variables. The following example shows the interest in this type of control with regard to safety. As for selective control, a set of rules is provided by the user to define the strategy to be implemented.

Example 6.5: **Pressure Control of a Chemical Gas Phase Reactor**

The reactor comprises only one measurement: pressure in the reactor. On the other hand, two valves must be actuated: one situated before the reactor, the other one after it. The actions on both valves must then be coordinated (Fig. 6.11).

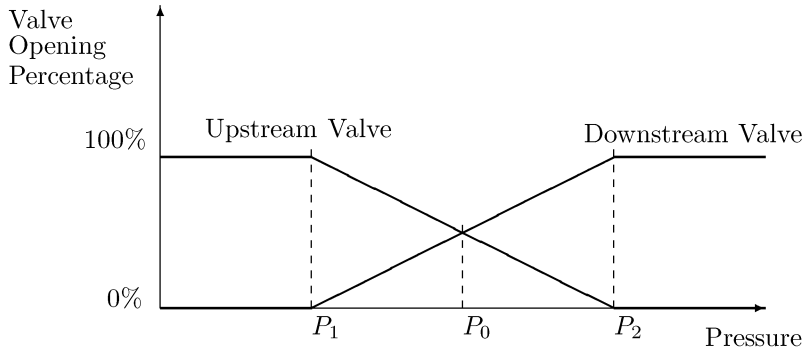


Figure 6.11: Split-range control

If the pressure is too high, the upstream valve is progressively closed according to a linear function of the deviation between the effective pressure and the desired pressure P_0, and the downstream valve is progressively opened. If the pressure is too low, the downstream valve is partially closed and the upstream valve opened. Below some fixed pressure P_1, the downstream valve is completely closed, and above a pressure P_2, the upstream valve is closed.

6.6 Feedforward Control

6.6.1 Generalities

Feedback control can never be perfect, as the manipulated variable is modified only when the influence of a disturbance is detected.

Feedforward control is a solution to this problem; the idea is to measure the disturbances and to take a decision before their effect is felt on the process (Fig. 6.12).

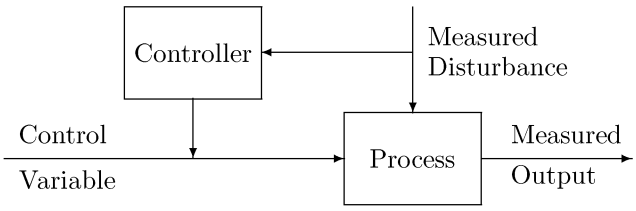


Figure 6.12: Feedforward control

Example 6.6: **Temperature and Concentration Control of a Perfectly Stirred Continuous Reactor**

Suppose that the only two disturbances are the concentration and the temperature of the feed stream. Both manipulated variables are the product withdrawal flow rate and the thermal fluid flow rate.

In classical feedback control, the temperature and concentration would have been measured in the reactor or in the outlet stream and the action would have been taken through both manipulated variables.

In feedforward control, the temperature and concentration of the inlet stream which constitute the disturbances are measured, and a model of the process is used to deduce the action on the same manipulated variables as previously.

Feedforward control suffers from several drawbacks:

- Disturbances must be measured on-line; this is not always possible.

- A model is necessary to know the manner according to which the controlled variables depend on the manipulated variables and on disturbances.
- An ideal feedforward controller, i.e. one that is able to maintain the process output to the set point value whatever the set point variations and the disturbances, is not always physically realizable.

6.6.2 Application in Distillation

Example 6.7: **Feedforward Control for a Distillation Column**
In a distillation column, frequently the feed is not controlled with respect to its

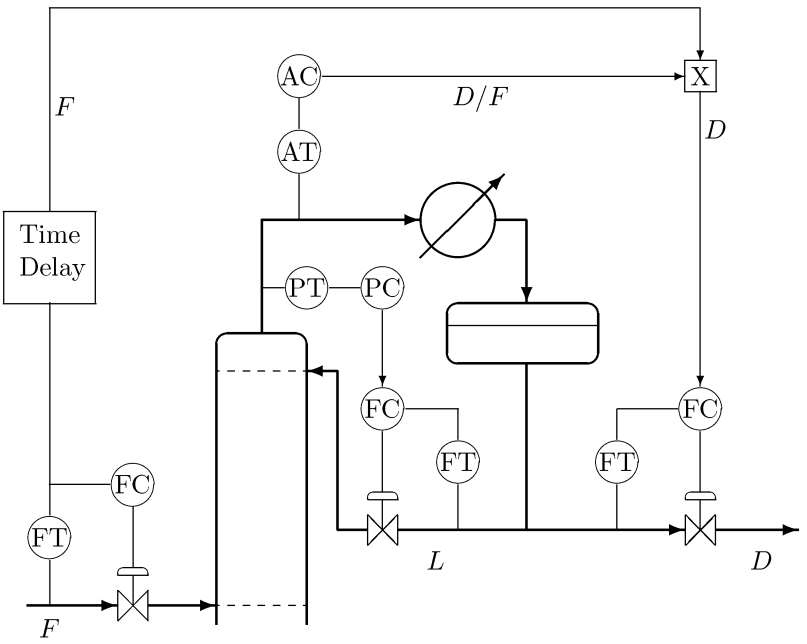


Figure 6.13: Realization of a feedforward control to control the feed flow rate disturbances

flow rate F and its composition z_F, so that, regarding control, these variables are considered as disturbances. Figure 6.13 shows how flow rate variations can be taken into account in a feedforward control (Ryskamp, 1987). The measurement of top pressure is used in cascade control: the output of the pressure controller PC is sent as a set point to the reflux (L) flow rate controller FC. The output of the top composition controller AC provides an estimation of the ratio D/F, which is multiplied by the measured flow rate F and, taking into account the time delay, gives the set point of the distillate D flow rate to the controller FC of the distillate flow rate.

In the proposed configuration, only the flow rate is measured, while the disturbances of feed composition are not measured. It would be possible to

build an additional feedforward control to take them into account.

6.6.3 Synthesis of a Feedforward Controller

A model of the process is necessary; the more accurate this model, the better the action of the feedforward controller.

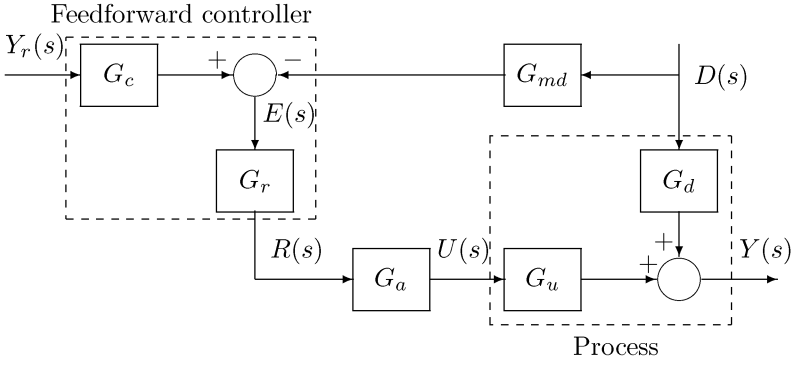


Figure 6.14: Block diagram of feedforward control

The process model is described by two transfer functions: one for the manipulated variable, the other one for the disturbance

$$Y(s) = G_u(s)\,U(s) + G_d(s)\,D(s) \tag{6.24}$$

To obtain perfect tracking, the following condition must be fulfilled

$$y(t) = y_r(t) \Longleftrightarrow Y(s) = Y_r(s) \tag{6.25}$$

so that ideally the following can be written

$$Y_r(s) = G_u(s)\,U(s) + G_d(s)\,D(s) \tag{6.26}$$

from which the input $U(s)$ is deduced, which takes into account the changes of the set point y_r and of the disturbance d

$$U(s) = \frac{1}{G_u(s)}\,Y_r(s) - \frac{G_d(s)}{G_u(s)}\,D(s) \tag{6.27}$$

To constitute a feedforward controller, the control system must include both terms of the set point and the disturbance and must also comprise two blocks. Thus, in the block diagram in Figure 6.14 (leave aside the measurement and actuator blocks) the feedforward control system includes, besides the comparator, a first block representing the main transfer function (related to the disturbance) of the feedforward controller equal to

$$\frac{U(s)}{D(s)} = -\frac{G_d(s)}{G_u(s)} \tag{6.28}$$

and a second block, the transfer function of which related to the set point is equal to $G_c(s)$ such as

$$\frac{U(s)}{Y_r(s)} = \frac{1}{G_u(s)} \tag{6.29}$$

From the block diagram, the control law integrating the feedforward controller is now written as

$$U(s) = G_c(s)\, G_r(s)\, G_a(s)\, Y_r(s) - G_{md}(s)\, G_r(s)\, G_a(s)\, D(s) \tag{6.30}$$

so that by identifying Eqs. (6.27) and (6.30), the following relations are obtained

$$\frac{1}{G_u(s)} = G_c(s)\, G_r(s)\, G_a(s) \qquad ; \qquad \frac{G_d(s)}{G_u(s)} = G_{md}(s)\, G_r(s)\, G_a(s) \tag{6.31}$$

giving

$$G_c(s) = \frac{G_{md}(s)}{G_d(s)} \tag{6.32}$$

and

$$G_r(s) = \frac{G_d(s)}{G_{md}(s)\, G_a(s)\, G_u(s)} \tag{6.33}$$

Checking:
Taking into account the disturbance measurement and the actuator, from the block diagram thus constituted, it can be written

$$Y(s)=[\, G_d(s) - G_{md}(s)\, G_r(s)\, G_a(s)\, G_u(s)\,]\; D(s)+G_c(s)\, G_r(s)\, G_a(s)\, G_u(s)\, Y_r(s) \tag{6.34}$$

1) So that the disturbance influence is zero (disturbance rejection), the factor of $D(s)$ must be zero, thus

$$G_r(s) = \frac{G_d(s)}{G_{md}(s)\, G_a(s)\, G_u(s)} \tag{6.35}$$

which results in

$$\begin{aligned} Y(s) &= G_c(s)\, G_r(s)\, G_a(s)\, G_u(s)\, Y_r(s) \\ &= \frac{G_c(s)\, G_d(s)}{G_{md}(s)} Y_r(s) \end{aligned} \tag{6.36}$$

2) A perfect set point tracking is desired, thus $Y(s) = Y_r(s)$, giving

$$G_c(s)\, G_d(s) = G_{md}(s) \tag{6.37}$$

hence

$$G_c(s) = \frac{G_{md}(s)}{G_d(s)} \tag{6.38}$$

Notice that the transfer functions thus calculated of the feedforward controller $G_r(s)$ and $G_c(s)$ depend on the process model through the reciprocals of the

transfer functions $G_d(s)$ and $G_u(s)$, which constitutes a main drawback of this control. To be physically realizable, a transfer function must have its numerator degree lower or equal to that of the denominator. Thus, when the theoretical transfer functions $G_r(s)$ and $G_c(s)$ do not fulfil this condition, it will be necessary to filter independently each of them by a filter of the form

$$\frac{1}{(\tau\, s + 1)^n} \tag{6.39}$$

Moreover, we must consider possible time delays or positive zeros present in the transfer functions intervening in the denominators of the expressions of $G_c(s)$ and $G_r(s)$, as they would correspond respectively to advances or unstable poles for $G_c(s)$ and $G_r(s)$. Consequently they must be eliminated in the calculation of $G_c(s)$ and $G_r(s)$.

6.6.4 Realization of a Feedforward Controller

This type of controller is, in fact, difficult to implement by an analog design; on the contrary, numerically, it is implementable on a microprocessor or a computer.

In the following discussion, to simplify, we consider that the transfer functions of the measurement and the actuator have no dynamics and thus are pure gains $G_m = K_m$ and $G_a = K_a$. If it were not the case, the following discussion should take the dynamics into account, but the reasoning would be similar.

Both transfer functions characteristic of the feedforward controller now depend, from a dynamic point of view, only on G_d and G_u. For these two transfer functions, two parts are distinguished: a steady-state element corresponding to the steady-state gain and a dynamic element, giving the notation

$$G_d(s) = K_d\, G'_d(s) \quad \text{and} \quad G_u(s) = K_u\, G'_u(s) \tag{6.40}$$

so that the steady-state gains of the dynamic parts $G'_d(s)$ and $G'_u(s)$ are equal to 1.

The feedforward controller itself will present two components

- A steady-state component.
- A dynamic component.

Calculation of the steady-state part:
It suffices to take for each transfer function its steady-state component, which gives the steady-state elements of the feedforward controller according to previous equations

$$K_r = \frac{K_d}{K_{md}\, K_a\, K_u} \tag{6.41}$$

and

$$K_c = \frac{K_{md}}{K_d} \tag{6.42}$$

Calculation of the dynamic part:

Rather than use the true transfer functions, first-order approximations are taken for $G'_d(s)$ and $G'_u(s)$, thus

$$G'_d(s) = \frac{1}{\tau_1\, s + 1} \quad \text{and} \quad G'_u(s) = \frac{1}{\tau_2\, s + 1} \tag{6.43}$$

The dynamic elements of the feedforward controller are then

$$G'_r(s) = \frac{G'_d(s)}{G'_u(s)} = \frac{\tau_2\, s + 1}{\tau_1\, s + 1} \tag{6.44}$$

and

$$G'_c(s) = \frac{1}{G'_d(s)} = \tau_1\, s + 1 \tag{6.45}$$

The dynamic part $G'_r(s)$ of the transfer function of the feedforward controller is called a lead-lag element, as the numerator $\tau_2\, s+1$ introduces a phase advance, while the denominator $\tau_1\, s + 1$ introduces a phase lag. According to the value of parameters τ_1 and τ_2, the lead-lag element can be either an advance or a lag. The theoretical transfer function G'_c is not physically realizable, so that it is necessary to filter it by at least a first-order filter of unit gain.

Physical Realizability and Pure Time Delays
Suppose that both transfer functions G'_d and G'_u contain respective time delays t_d and t_u and that the rest of the transfer function is constituted by the same first-order

$$G'_d(s) = \frac{\exp(-t_d\, s)}{\tau\, s + 1} \quad \text{and} \quad G'_u(s) = \frac{\exp(-t_u\, s)}{\tau\, s + 1} \tag{6.46}$$

In this case, it would result in

$$G'_r = \frac{G'_d}{G'_u} = \exp((t_u - t_d)\, s) \tag{6.47}$$

which means that if t_u is larger than t_d, then this is an advance term! Thus, future values of the disturbance would have to be known to calculate the value of the input. This is, of course, impossible: the ideal feedforward controller is physically unrealizable. A physically realizable controller can be constituted by an approximation of the theoretical controller.

6.6.5 Feedforward and Feedback Control

As the requirements of feedforward control are very important, it seems interesting to add to it a feedback control which will correct its imperfections (Fig. 6.15):

- Non identified or ill-known disturbances.
- Variations in the operating conditions of the process.
- Model of the incomplete process.

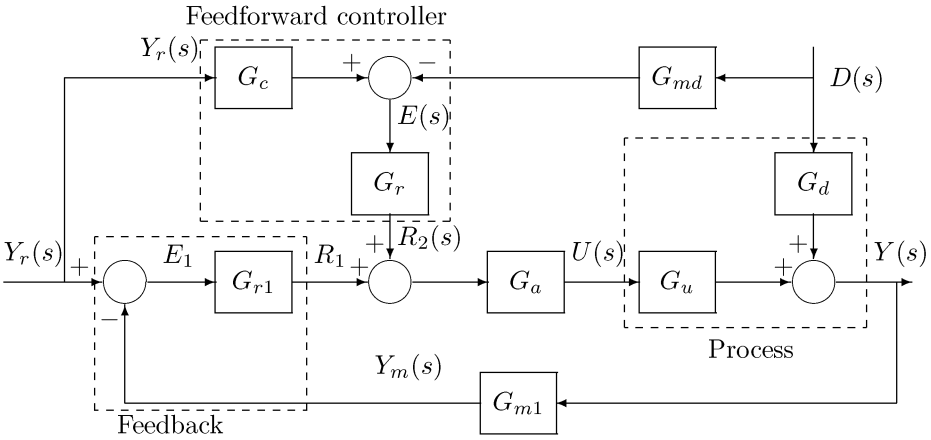


Figure 6.15: Block diagram of feedforward and feedback control

However, the philosophy is that the feedforward controller should do 90% of the job while the feedback would correct the remaining 10%.
According to the block diagram (Fig. 6.15), the expression of the system output Y with respect to inputs Y_r and D from the system equations

$$Y = G_u\, U + G_d\, D \tag{6.48}$$

and

$$U = G_a\, (U_{a1} + U_{a2}) = G_a\, (G_{r1}\, E_1 + G_{r2}\, E_2) \tag{6.49}$$

with

$$E_2 = G_c\, Y_r - G_{md}\, D \tag{6.50}$$

and

$$E_1 = Y_r - Y_m = Y_r - G_{m1}\, Y \tag{6.51}$$

thus

$$Y = G_u\, G_a\, G_{r1}\, (Y_r - G_{m1}\, Y) + G_u\, G_a\, G_{r2}\, (G_c\, Y_r - G_{md}\, D) + G_d\, D \tag{6.52}$$

hence

$$Y = \frac{G_u\, G_a\, (G_{r1} + G_{r2}\, G_c)}{1 + G_u\, G_a\, G_{r1}\, G_{m1}} Y_r + \frac{G_d - G_u\, G_a\, G_{r2}\, G_{md}}{1 + G_u\, G_a\, G_{r1}\, G_{m1}} D \tag{6.53}$$

The stability analysis realized for feedback systems is not modified by the addition of a feedforward loop. Actually, the characteristic equation:
$1 + G_u\, G_a\, G_{r1}\, G_{m1} = 0$ depends only on the feedback elements.

To find the characteristics of the feedforward system, the reasoning is based on the fact that the control work must be realized by the feedforward action. A perfect disturbance rejection is demanded, thus

$$G_d = G_u\, G_a\, G_{r2}\, G_{md} \tag{6.54}$$

hence

$$G_{r2} = \frac{G_d}{G_u\, G_a\, G_{md}} \tag{6.55}$$

To find the transfer function linked to the set point, the set point Y_r can be separated into two terms: one Y_{r1} acting on the feedback and the other one Y_{r2} acting on the feedforward. Thus, we can consider $Y_{r1} = 0$ and $D = 0$,
This results in

$$G_u\, G_a\, G_{r2}\, G_c = 1 \tag{6.56}$$

hence

$$G_c = \frac{G_{md}}{G_d} \tag{6.57}$$

If the feedforward action were not perfect, a deviation $e_2 \neq 0$ would occur, and the feedback would partially compensate the deviation.

6.7 Ratio Control

In ratio control, the objective is to maintain two measured variables at a constant ratio. This control is most often applied in the case of flow rates ratios. The remarkable characteristic is that only one of the flow rates is manipulated, denoted by m; the other flow rate is called "wild" and corresponds to a disturbance d. A typical application is the control of the reflux rate of a distillation column: the ratio of the reflux flow rate to the distillate flow rate must be maintained constant.

The desired ratio R is thus equal to

$$R = \frac{m}{d} \tag{6.58}$$

This control can be realized in one of two ways:

- Either both flow rates ("wild" and manipulated) are measured and their calculated ratio is compared to a desired ratio; the deviation between the ratios is used as the input of the ratio controller to correct the manipulated stream. This method implies the use of a divider (nonlinear). The process open-loop gain in this case is equal to

$$K = \left(\frac{\partial R}{\partial m}\right)_d = \frac{1}{d} \tag{6.59}$$

  and thus is the inverse of the disturbance ("wild" stream).
- Either the "wild" flow rate is measured and multiplied by the desired ratio, then compared to the flow rate that the other manipulated stream should have. In this method, the process open-loop gain remains constant.

In ratio control, it appears that in both cases the disturbance must be measured. Thus, it can be considered as a particular form of feedforward control.

Example 6.8: **Ratio Control for a Distillation Column**

For a distillation column, Ryskamp (1980) proposed to manipulate the ratio

D/V_n to control the top composition. Because of the mass balance $V_n = L+D$, if D/V_n is fixed, L/D and L/V_n are also fixed. The system described by Fig. 6.16 responds to heat power variations at the reboiler in the following manner:

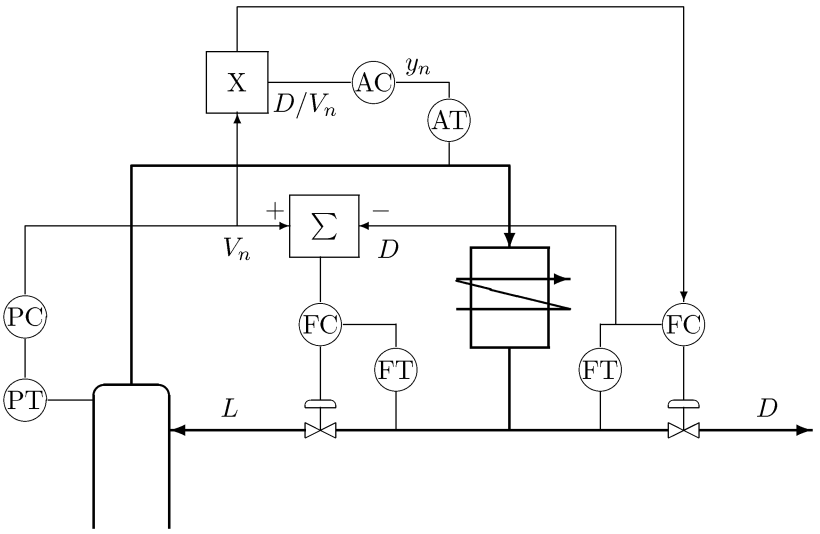


Figure 6.16: Ratio control at the top of a distillation column

increasing the heat power induces a pressure increase, which is counterbalanced by an increase of its output, i.e. the total top vapour V_n. This signal is on one hand multiplied by the output D/V_n of the composition controller to provide the set point of distillate D flow rate. On the other hand, it enters the comparator, where the measured flow rate D is subtracted from it to give the set point of the reflux L flow rate. The pressure controller thus increases L and D according to a ratio set by the output of the composition controller. Thus, the reflux rate varies without a time delay at the request of composition controller.

References

S.I. Niculescu and A.M Annaswamy. An adaptive Smith-controller for time-delay systems with relative degree $n \leq 2$. pages 1–15, 2003.

C.J. Ryskamp. New strategy improves dual composition control. *Hydrocarbon processing*, (6), 1980.

C.J. Ryskamp. Dual composition column control. In Les Kane, editor, *Handbook of Advanced Process Control Systems and Instrumentation*, pages 158–. Gulf Publishing Company, Houston, 1987.

J.R. Schleck and D. Danesian. An evaluation of the Smith linear predictor technique for controlling deadtime dominated processes. *ISA Trans*, 17(4): 39–, 1978.

O.J.M. Smith. Close control of loops with dead time. *Chem. Eng. Prog.*, 53 (5):217–, 1957.

O.J.M. Smith. A controller to overcome deadtime. *ISA Journal*, 6(2):28–33, 1959.

V.J. VanDoren. The Smith predictor: a process engineer's crystal ball. *Control Engineering*, May 1, 1996.

Chapter 7

State Representation, Controllability and Observability

In this chapter, only notions relative to state-space linear systems are studied. The fundamental concepts of controllability and observability are described in an analogous manner in continuous and discrete time; they even can be defined independently of the state-space representation.

7.1 State Representation

7.1.1 Monovariable System

The dynamic behaviour of a monovariable system (single-input single-output) is expressed by an ordinary differential equation of order n (equivalent to a set of n first-order ordinary differential equations) or a partial differential equation; for a multivariable system, it is transformed into a set of ordinary differential equations. Using a numerical scheme of discretization of ordinary derivatives, it is possible to represent any system under a discrete form in the state space.

A continuous monovariable system is described by the following state-space representation

$$\begin{array}{ll} \dot{\boldsymbol{x}}(t) & = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}u(t) \\ y(t) & = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}u(t) \end{array} \tag{7.1}$$

$\boldsymbol{x}$ is the state vector of dimension n. $\boldsymbol{A}(n \times n)$ is the state matrix (or evolution matrix), $\boldsymbol{B}(n \times 1)$ the input matrix (or control matrix) and $\boldsymbol{C}(1 \times n)$ the output matrix (or observation matrix). D is a scalar in the monovariable case.

The set of differential equations

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}u(t) \qquad t \geq 0 \tag{7.2}$$

can be integrated between 0 and t when the matrix $\boldsymbol{A}$ has constant coefficients, and its solution $\boldsymbol{x}(t, \boldsymbol{x}(0), u)$ is

$$\boldsymbol{x}(t, \boldsymbol{x}(0), u) = \exp(\boldsymbol{A}t)\boldsymbol{x}(0) + \int_0^t \exp(\boldsymbol{A}(t-\tau))\boldsymbol{B}u(\tau)d\tau \tag{7.3}$$

When the coefficients of the matrix $\boldsymbol{A}$ are not constant, the solution corresponding to an integration between two instants t and t_0 is represented in the form

$$\boldsymbol{x}(t, \boldsymbol{x}(t_0), u) = \boldsymbol{\Phi}(t, t_0)\boldsymbol{x}(t_0) + \int_{t_0}^t \boldsymbol{\Phi}(t, \tau)\boldsymbol{B}u(\tau)d\tau \tag{7.4}$$

where $\boldsymbol{\Phi}(t, t_0)$ is called the state transition matrix corresponding to the integration of the homogeneous equation

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}(t)\boldsymbol{x}(t) \tag{7.5}$$

The monovariable system can be represented by the following continuous transfer function

$$G(s) = \boldsymbol{C}(s\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B} + D \tag{7.6}$$

The state representation $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, D)$ constitutes a realization of $G(s)$ and is not unique. It suffices to apply an invertible transformation matrix $\boldsymbol{T}$ on this realization to find a different state representation

$$\boldsymbol{A}' = \boldsymbol{T}^{-1}\,\boldsymbol{A}\,\boldsymbol{T} \quad , \quad \boldsymbol{B}' = \boldsymbol{T}^{-1}\,\boldsymbol{B} \quad , \quad \boldsymbol{C}' = \boldsymbol{C}\,\boldsymbol{T} \tag{7.7}$$

To the continuous monovariable system, corresponds, by time discretization, the following discrete system

$$\begin{array}{ll} \boldsymbol{x}(k+1) & = \boldsymbol{A}_d\boldsymbol{x}(k) + \boldsymbol{B}_d u(k) \\ y(k) & = \boldsymbol{C}\boldsymbol{x}(k) + \boldsymbol{D}u(k) \end{array} \tag{7.8}$$

where the instant t_k is simply indicated by integer variable k.

When sampling is performed with a period T_s and a zero-order holder, it results that

$$\begin{array}{rcl} \boldsymbol{A}_d & = & \exp(\boldsymbol{A}T_s) \\ \boldsymbol{B}_d & = & \displaystyle\int_0^{T_s} \exp(\boldsymbol{A}t)\boldsymbol{B}dt \end{array} \tag{7.9}$$

The monovariable system (7.8) can be represented by the following discrete transfer function

$$H(z) = \boldsymbol{C}(z\boldsymbol{I} - \boldsymbol{A}_d)^{-1}\boldsymbol{B}_d + D \tag{7.10}$$

To proceed from the continuous transfer function $G(s)$ of Eq. (7.6) to the discrete transfer function of Eq. (7.10), it thus suffices to consider s or z as operators and to replace $\boldsymbol{A}$ and $\boldsymbol{B}$ by $\boldsymbol{A}_d$ and $\boldsymbol{B}_d$.

7.1.2 Multivariable System

The multivariable (multi-input multi-output) state representation is very close to the single-input single-output state representation. The dynamic behaviour of a monovariable system is expressed by an ordinary differential equation of order n or partial differential equation; for a multivariable system, it is a set of differential equations. A continuous multivariable system can be represented by the model

$$\begin{array}{ll} \dot{\boldsymbol{x}}(t) & = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \\ \boldsymbol{y}(t) & = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) \end{array} \tag{7.11}$$

If the state vector $\boldsymbol{x}(t)$ has dimension n, $\boldsymbol{A}$ state matrix (or evolution matrix) has dimension $(n \times n)$, $\boldsymbol{B}$ input matrix (or control matrix) has dimension $(n \times n_u)$ $\boldsymbol{C}$ output matrix (or observation matrix) has dimension $(n_y \times n)$, while $\boldsymbol{D}$ is the coupling matrix (or direct transmission matrix) of dimension $(n_y \times n_u)$.

These equations can be transformed to obtain the system with respect to the Laplace variable

$$\begin{array}{lcl} s\,\boldsymbol{X}(s) & = & \boldsymbol{A}\,\boldsymbol{X}(s) + \boldsymbol{B}\,\boldsymbol{U}(s) \\ \boldsymbol{Y}(s) & = & \boldsymbol{C}\,\boldsymbol{X}(s) + \boldsymbol{D}\,\boldsymbol{U}(s) \end{array} \tag{7.12}$$

The Laplace transform of the state vector is

$$\boldsymbol{X}(s) = [s\,\boldsymbol{I} - \boldsymbol{A}]^{-1}\,\boldsymbol{B}\,\boldsymbol{U}(s) \tag{7.13}$$

and the Laplace transform of the output is equal to

$$\boldsymbol{Y}(s) = \boldsymbol{C}\,[s\,\boldsymbol{I} - \boldsymbol{A}]^{-1}\,\boldsymbol{B}\,\boldsymbol{U}(s) + \boldsymbol{D}\,\boldsymbol{U}(s) \tag{7.14}$$

The multivariable system (7.11) can be represented by the matrix of continuous transfer functions of the form

$$\boldsymbol{G}(s) = \boldsymbol{C}(s\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B} + \boldsymbol{D} \tag{7.15}$$

For sufficiently large $|s|$, $G(s)$ can be expressed as a Markov series expansion

$$\boldsymbol{G}(s) = \boldsymbol{D} + \frac{\boldsymbol{C}\boldsymbol{B}}{s} + \frac{\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}}{s^2} + \frac{\boldsymbol{C}\boldsymbol{A}^2\boldsymbol{B}}{s^3} + \ldots \tag{7.16}$$

where the factors $\boldsymbol{C}\boldsymbol{A}^{i-1}\boldsymbol{B}$ are Markov parameters.

By using a numerical scheme of discretization of ordinary derivatives, any system can be represented in a discrete form in the state space. The discrete state-space model is

$$\begin{array}{ll} \boldsymbol{x}(t+1) & = \boldsymbol{A}_d\boldsymbol{x}(t) + \boldsymbol{B}_d\boldsymbol{u}(t) \\ \boldsymbol{y}(t) & = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) \end{array} \tag{7.17}$$

The state vector $\boldsymbol{x}(t)$ has dimension n, $\boldsymbol{A}_d$ has dimension $(n \times n)$, $\boldsymbol{B}_d$ dimension $(n \times n_u)$, $\boldsymbol{C}$ dimension $(n_y \times n)$ and $\boldsymbol{D}$ dimension $(n_y \times n_u)$.

This multivariable system can be represented by the matrix of discrete transfer functions in the form

$$\boldsymbol{H}(z) = \boldsymbol{C}(z\boldsymbol{I} - \boldsymbol{A}_d)^{-1}\boldsymbol{B}_d + \boldsymbol{D} \tag{7.18}$$

7.2 Controllability

The reasoning concerning controllability, observability and realizations, is identical for a continuous-time and a discrete-time system. Although examples will be chosen for single-input single-output systems, controllability and observability concepts can be easily generalized for multivariable systems.

Equation (7.17) shows that to generate outputs $y_j(k)$ from inputs $u_i(k)$, the states $\boldsymbol{x}(k)$, which themselves depend on inputs $u_i(k)$, are necessary intermediates, thus these states must be controllable. Moreover, it is important that the states $\boldsymbol{x}(k)$ can be reconstructed from the measurements constituted by inputs $u_i(k)$ and outputs $y_j(k)$, thus the states must be observable.

A system is said to be controllable if it is possible to find an input $\boldsymbol{u}$ that allows it to go from a state $\boldsymbol{x}(1)$ to a state $\boldsymbol{x}(2)$ in a finite time. Often, the state $\boldsymbol{x}(1)$ is the initial state and time t_1 the initial instant.

This condition is not necessarily realized. Consider a monovariable system: it is sufficient that one or several states is not influenced by the input.

Some authors such as Wohnam (1985) distinguish reachability from controllability. According to Wohnam (1985), a state $\boldsymbol{x}$ is reachable from a state $\boldsymbol{x}(0)$ if there exists a time t $(0 < t < \infty)$ and an input $\boldsymbol{u}$ (in the domain of acceptable inputs) such that the solution $\boldsymbol{x}(t, \boldsymbol{x}(0), \boldsymbol{u})$ of the linear ordinary differential Eq. (7.2) is equal to this given $\boldsymbol{x}$. Let $\mathcal{A}$ be the space of reachable states from the state $\boldsymbol{x}(0)$ and $\mathcal{X}$ the space of all states; $\mathcal{A}$ is a subspace of $\mathcal{X}$. The linear system described by the pair $(\boldsymbol{A}, \boldsymbol{B})$ is controllable if the subspace $\mathcal{A}$ is the entire space $\mathcal{X}$ (their dimension is the same). In the following, only controllability properties will be studied.

In the case of a discrete monovariable system, Eq. (7.8) allows us to write at successive instants, using recurrent relations

$$\begin{aligned}
\boldsymbol{x}(1) &= \boldsymbol{A}\boldsymbol{x}(0) + \boldsymbol{B}u(0) \\
\boldsymbol{x}(2) &= \boldsymbol{A}\boldsymbol{x}(1) + \boldsymbol{B}u(1) \quad = \boldsymbol{A}^2\boldsymbol{x}(0) + [\boldsymbol{A}\,\boldsymbol{B}u(0) + \boldsymbol{B}u(1)] \\
\boldsymbol{x}(3) &= \boldsymbol{A}\boldsymbol{x}(2) + \boldsymbol{B}u(2) \quad = \boldsymbol{A}^3\boldsymbol{x}(0) + [\boldsymbol{A}^2\,\boldsymbol{B}u(0) + \boldsymbol{A}\,\boldsymbol{B}u(1) + \boldsymbol{B}u(2)] \\
&\vdots \\
\boldsymbol{x}(k) &= \boldsymbol{A}\boldsymbol{x}(k-1) + \boldsymbol{B}u(k-1) \\
&= \boldsymbol{A}^k\boldsymbol{x}(0) + [\boldsymbol{A}^{k-1}\,\boldsymbol{B}u(0) + \boldsymbol{A}^{k-2}\,\boldsymbol{B}u(1) + \ldots + \boldsymbol{B}u(k-1)]
\end{aligned} \tag{7.19}$$

The state vector, being of dimension n, for $k = n$, Eq. (7.19) is written as

$$\boldsymbol{x}(n) - \boldsymbol{A}^n\boldsymbol{x}(0) = [\boldsymbol{B} \quad \boldsymbol{A}\,\boldsymbol{B} \quad \ldots \quad \boldsymbol{A}^{n-2}\,\boldsymbol{B} \quad \boldsymbol{A}^{n-1}\,\boldsymbol{B}] \begin{bmatrix} u(n-1) \\ u(n-2) \\ \vdots \\ u(1) \\ u(0) \end{bmatrix} \tag{7.20}$$

Representing by $\boldsymbol{u}_{n-1}$ the vector composed by successive inputs, and starting from the most recent inputs

$$\boldsymbol{u}_{n-1}^T = [u(n-1) \quad u(n-2) \quad \ldots \quad u(1) \quad u(0)] \tag{7.21}$$

the necessary and sufficient condition for the vector of inputs $\boldsymbol{u}_{n-1}$ to be calculable is that the controllability matrix

$$\boldsymbol{\mathcal{C}} = [\boldsymbol{B} \quad \boldsymbol{A}\,\boldsymbol{B} \quad \boldsymbol{A}^{n-2}\,\boldsymbol{B} \quad \boldsymbol{A}^{n-1}\,\boldsymbol{B}] \tag{7.22}$$

has rank n. The pair $(\boldsymbol{A}, \boldsymbol{B})$ is said to be controllable.

At any instant $k > n$, the rank of matrix

$$\boldsymbol{\mathcal{C}}_k = [\boldsymbol{B} \quad \boldsymbol{A}\,\boldsymbol{B} \quad \boldsymbol{A}^{k-2}\,\boldsymbol{B} \quad \boldsymbol{A}^{k-1}\,\boldsymbol{B}] \tag{7.23}$$

would have been the same as that of matrix $\boldsymbol{\mathcal{C}}$.

The necessary and sufficient condition of controllability of a system is thus that the controllability matrix $\boldsymbol{\mathcal{C}}$ has full rank and is equal to n.

The rank condition, which is mathematically strict, will only be approached at the numerical level[1].

The system described by the linear difference equation

$$y(k) + a_1\,y(k-1) + \ldots + a_n\,y(k-n) = b_1\,u(k-1) + \ldots + b_n\,u(k-n) \tag{7.24}$$

corresponding to the discrete transfer function

$$G(z) = \frac{b_1 z^{-1} + \ldots + b_n z^{-n}}{1 + a_1 z^{-1} + \ldots + a_n z^{-n}} \tag{7.25}$$

can be described by the following controllable canonical form

$$\boldsymbol{A}_c = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & 1 \\ -a_n & -a_{n-1} & \ldots & \ldots & -a_1 \end{bmatrix} \qquad \boldsymbol{B}_c = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{7.26}$$

$$\boldsymbol{C}_c = \begin{bmatrix} b_n & b_{n-1} & \ldots & b_2 & b_1 \end{bmatrix} \tag{7.27}$$

where the matrices $(\boldsymbol{A}_c, \boldsymbol{B}_c, \boldsymbol{C}_c)$ are expressed in the controllable companion form.

The pair $(\boldsymbol{A}_c, \boldsymbol{B}_c)$ such defined is always controllable; to prove it, the corresponding controllability matrix can be shown to be of rank n. The matrix $\boldsymbol{A}_c$ is a companion form of the denominator of the transfer function.

Note that there exist four canonical controllable forms (Borne et al., 1992; Foulard et al., 1987) that are possible for a monovariable system, depending on whether the sequence $\{a_1 \ \ldots \ a_n\}$ is situated on the first or the last row, the first or the last column of matrix $\boldsymbol{A}_c$. For this reason, the resulting matrices $\boldsymbol{B}_c$ and $\boldsymbol{C}_c$ differ. Note that the matrix $\boldsymbol{A}$ from the canonical controllable

[1]The rank of a matrix $\boldsymbol{A}$ can be numerically calculated as the number of singular values of $\boldsymbol{A}$ that are larger than a given scalar $\epsilon > 0$.

or observable forms is often ill-conditioned [2] and thus will make numerical calculation difficult and inaccurate.

Consider $(\boldsymbol{A}_0, \boldsymbol{B}_0, \boldsymbol{C}_0)$ as a given controllable form. Let $\boldsymbol{P}$ be a transformation matrix such that $\boldsymbol{A} = \boldsymbol{P}\,\boldsymbol{A}_0\,\boldsymbol{P}^{-1}$, $\boldsymbol{B} = \boldsymbol{P}\,\boldsymbol{B}_0$ and $\boldsymbol{C} = \boldsymbol{C}_0\,\boldsymbol{P}^{-1}$. It can be shown that $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ constitutes a new controllable form. Thus, again the non-uniqueness of the state-space representation of a given system is shown.

If a state-space system is considered in any form $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, the polynomial characteristic of matrix $\boldsymbol{A}$ equal to

$$\det(s\boldsymbol{I} - \boldsymbol{A}) = s^n + a_1\, s^{n-1} + \ldots + a_{n-1}\, s + a_n \tag{7.28}$$

depends only on the coefficients of the last line of matrix $\boldsymbol{A}_c$ of its controllable canonical form.

The state-space system can be represented by using the eigenvalues of matrix $\boldsymbol{A}$ or modes of the system. The form thus obtained is called the modal canonical form or Jordan form which will be denoted by $(\boldsymbol{A}_{mc}, \boldsymbol{B}_{mc}, \boldsymbol{C}_{mc})$:

- When the eigenvalues of matrix $\boldsymbol{A}$ are real, they constitute the diagonal elements of matrix $\boldsymbol{A}_{mc}$, or

$$\boldsymbol{A}_{mc} = \begin{bmatrix} \lambda_1 & 0 & \ldots \\ 0 & \ddots & 0 \\ \vdots & \ldots & \lambda_n \end{bmatrix} \tag{7.29}$$

- When the eigenvalues of matrix $\boldsymbol{A}$ are complex, they constitute diagonal blocks of matrix $\boldsymbol{A}_{mc}$. Let: $\lambda_1 = \sigma + j\omega$ be a complex eigenvalue, its conjugate is also a complex eigenvalue. The diagonal block is

$$\begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix} \tag{7.30}$$

  and is thus formed by the real parts of the complex eigenvalues on the diagonal, and the complex parts on each side of the diagonal.
- When λ_1 is a multiple eigenvalue of order j, there corresponds in matrix $\boldsymbol{A}_{mc}$ a square block of dimension j of the form

$$\begin{bmatrix} \lambda & 1 & 0 & \ldots \\ 0 & \lambda & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ 0 & \ldots & 0 & \lambda \end{bmatrix} \tag{7.31}$$

The matrix $\boldsymbol{A}_{mc}$ corresponding to the modal canonical form is well-conditioned and it is recommended to use it rather than a companion form.

[2] The condition number of a matrix is the ratio of the largest singular value to the smallest singular value of this matrix. A matrix is well-conditioned and thus easily invertible when this value is around 1. When the condition number becomes larger than around 1000 or more, the matrix is increasingly ill-conditioned.

The matrix $\boldsymbol{B}_{mc}$ is only formed of 1, except for the $j-1$ first lines corresponding to a multiple pole of order j where these are 0.

The system transfer function corresponding to the modal canonical form corresponds to a decomposition as a sum of rational fractions where simple, multiple and conjugate complex poles are demonstrated. The coefficients of rational fractions are the elements of matrix $\boldsymbol{C}_{mc}$.

Example 7.1: **An Uncontrollable System**

Consider the example in Figure 7.1. The continuous state-space model of this

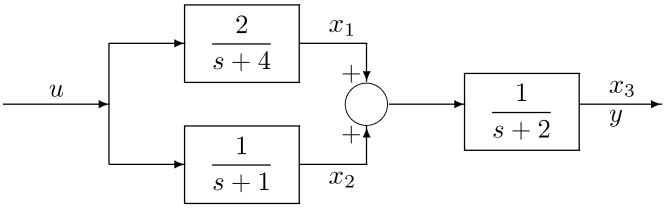


Figure 7.1: Uncontrollable system. System block diagram corresponding to Eq. (7.32)

system is given by

$$\boldsymbol{A} = \begin{bmatrix} -4 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 1 & -2 \end{bmatrix} \quad ; \quad \boldsymbol{B} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \quad ; \quad \boldsymbol{C} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \tag{7.32}$$

The controllability and observability matrices are respectively

$$\boldsymbol{\mathcal{C}} = \begin{bmatrix} 2 & -8 & 32 \\ 1 & -1 & 1 \\ 0 & 3 & -15 \end{bmatrix} \quad ; \quad \boldsymbol{\mathcal{O}} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & -2 \\ -6 & -3 & 4 \end{bmatrix} \tag{7.33}$$

The controllability matrix $\boldsymbol{\mathcal{C}}$ has rank 2 and the observability matrix $\boldsymbol{\mathcal{O}}$ has rank 3. Thus, one of the states of this system is not controllable.

The transfer function corresponding to this system is equal to

$$G(s) = \frac{3}{(s+4)(s+1)} \tag{7.34}$$

It is clear that the pole $s = -2$ has been cancelled.

The modal canonical form of this system is obtained by a basis change using a transformation matrix P. The modal canonical form thus obtained is

$$\boldsymbol{A}_{mc} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & -1 \end{bmatrix}; \boldsymbol{B}_{mc} = \begin{bmatrix} 0 \\ 2,2361 \\ 1,4142 \end{bmatrix}; \boldsymbol{C}_{mc} = \begin{bmatrix} 1 & -0,4472 & 0,7071 \end{bmatrix} \tag{7.35}$$

while the transformation matrix is

$$\boldsymbol{P} = \begin{bmatrix} 0.5 & -1 & 1 \\ 1.118 & 0 & 0 \\ 0 & 1.4142 & 0 \end{bmatrix} \tag{7.36}$$

The block representation of this system (Fig. 7.2) shows that the state x_1 corresponding to the eigenvalue -2 is not controllable.

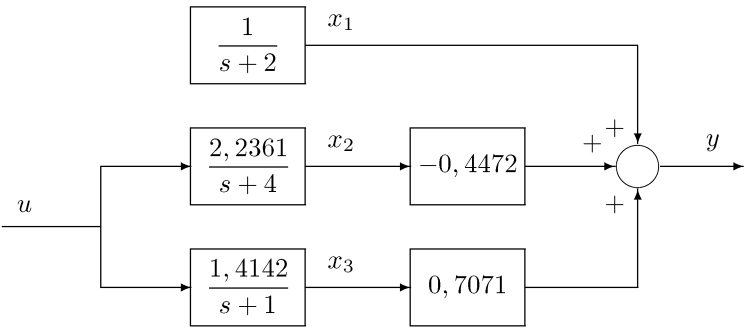


Figure 7.2: Uncontrollable system. Block diagram corresponding to the modal canonical form (Eq. (7.35))

It can be noticed that the matrix $\boldsymbol{A}$ already had its eigenvalues on the diagonal, as it was lower triangular.

7.3 Observability

A system is said to be observable if, in an interval $[t_1, t_2]$, with the input assumed to be known a priori, the knowledge of the output allows us to determine the initial state $x(t_1)$ of the system. It suffices that a state does not influence the output for this system not to be observable.

Consider the monovariable discrete system (7.8). The output at time k is equal to

$$y(k) = \boldsymbol{C}\boldsymbol{A}^k\boldsymbol{x}(0) + \boldsymbol{C}\boldsymbol{A}^{k-1}\boldsymbol{B}u(0) + \boldsymbol{C}\boldsymbol{A}^{k-2}\boldsymbol{B}u(1) + \ldots + \boldsymbol{C}\boldsymbol{B}u(k-1) + Du(k) \tag{7.37}$$

This recurrent relation can be used to make the state vector appear at the

initial instant with respect to all past inputs and outputs according to

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \dots \\ y(k) \end{bmatrix} = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{CA} \\ \boldsymbol{CA}^2 \\ \dots \\ \boldsymbol{CA}^k \end{bmatrix} \boldsymbol{x}(0) + \begin{bmatrix} D & 0 & \dots & \dots & 0 \\ \boldsymbol{CB} & D & 0 & \dots & 0 \\ \boldsymbol{CAB} & \boldsymbol{CB} & D & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \boldsymbol{CA}^{k-1}\boldsymbol{B} & \boldsymbol{CA}^{k-2}\boldsymbol{B} & \dots & \boldsymbol{CB} & D \end{bmatrix} \begin{bmatrix} u(0) \\ u(1) \\ u(2) \\ \dots \\ u(k-1) \end{bmatrix} \tag{7.38}$$

To calculate the initial state $\boldsymbol{x}(0)$ of dimension n with respect to past inputs and outputs, the matrix $\boldsymbol{\mathcal{O}}_k$

$$\boldsymbol{\mathcal{O}}_k = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{C\,A} \\ \dots \\ \boldsymbol{C\,A}^{k-1} \\ \boldsymbol{C\,A}^{k} \end{bmatrix} \tag{7.39}$$

should have rank n. If $k > n-1$, the rank of matrix $\boldsymbol{\mathcal{O}}_k$ is at maximum n.

The necessary and sufficient condition of observability is thus that the observability matrix $\boldsymbol{\mathcal{O}}$

$$\boldsymbol{\mathcal{O}} = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{C\,A} \\ \dots \\ \boldsymbol{C\,A}^{n-2} \\ \boldsymbol{C\,A}^{n-1} \end{bmatrix} \tag{7.40}$$

has full rank, equal to n. The pair $(\boldsymbol{A}, \boldsymbol{C})$ is said to be observable.

Again consider the system described by the difference linear Eq. (7.24). The system can be represented by the following observable canonical form

$$\boldsymbol{A}_o = \begin{bmatrix} 0 & \dots & \dots & 0 & -a_n \\ 1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & 1 & -a_1 \end{bmatrix} \qquad \boldsymbol{B}_o = \begin{bmatrix} b_n \\ \vdots \\ \vdots \\ \vdots \\ b_1 \end{bmatrix} \tag{7.41}$$

$$\boldsymbol{C}_o = \begin{bmatrix} 0 & \dots & \dots & 0 & 1 \end{bmatrix} \tag{7.42}$$

where the matrices $(\boldsymbol{A}_o, \boldsymbol{B}_o, \boldsymbol{C}_o)$ are expressed in the observable companion form.

The pair $(\boldsymbol{A}_o, \boldsymbol{C}_o)$ thus defined is always observable; to prove it, it suffices to show that the corresponding observability matrix has rank n. The previous matrix $\boldsymbol{A}_o$ is also a companion form of the denominator of the transfer function.

There exist also four possible canonical observable forms (Borne et al., 1992; Foulard et al., 1987), for a monovariable system.

In a similar manner to the controllability study, if a state-space system is considered under any form $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, the characteristic polynomial of the matrix $\boldsymbol{A}$ equal to

$$\det(s\boldsymbol{I} - \boldsymbol{A}) = s^n + a_1\, s^{n-1} + \ldots + a_{n-1}\, s + a_n \tag{7.43}$$

depends only on the coefficients of the last column of the matrix $\boldsymbol{A}_o$ of its observable canonical form.

Controllability and observability are dual notions. It is possible, for example, to go from the controllability matrix to the observability one by considering a second system where $\boldsymbol{A}$ and $\boldsymbol{B}$ are respectively replaced by $\boldsymbol{A}^T$ and $\boldsymbol{C}^T$.

Example 7.2: **An Unobservable System**

Consider the example in Figure 7.3. Notice the similarity with the example used

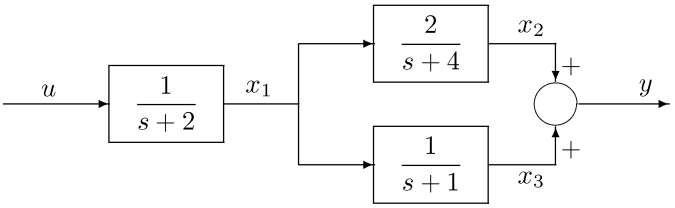


Figure 7.3: Unobservable system. Block diagram of the system corresponding to Eq. (7.44)

to illustrate the loss of controllability. The state-space model of this system is given by

$$\boldsymbol{A} = \begin{bmatrix} -2 & 0 & 0 \\ 2 & -4 & 0 \\ 1 & 0 & -1 \end{bmatrix} \quad ; \quad \boldsymbol{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad ; \quad \boldsymbol{C} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \tag{7.44}$$

The controllability and observability matrices are, respectively

$$\boldsymbol{\mathcal{C}} = \begin{bmatrix} 1 & -2 & 4 \\ 0 & 2 & -12 \\ 0 & 1 & -3 \end{bmatrix} \quad ; \quad \boldsymbol{\mathcal{O}} = \begin{bmatrix} 0 & 1 & 1 \\ 3 & -4 & -1 \\ -15 & 16 & 1 \end{bmatrix} \tag{7.45}$$

The controllability matrix $\boldsymbol{\mathcal{C}}$ has rank 3 and the observability matrix $\boldsymbol{\mathcal{O}}$ has rank 2. One of the states of this system is thus not observable.

The transfer function corresponding to this system is the same as in the previous example and is equal to

$$G(s) = \frac{3}{(s+4)(s+1)} \tag{7.46}$$

The pole $s = -2$ has been cancelled. The modal canonical form is

$$\boldsymbol{A}_{mc} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & -2 \end{bmatrix} \quad ; \quad \boldsymbol{B}_{mc} = \begin{bmatrix} 1 \\ -1 \\ 1.7321 \end{bmatrix} \quad ; \quad \boldsymbol{C}_{mc} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \tag{7.47}$$

while the transformation matrix is

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1.7321 & 0 & 0 \end{bmatrix} \tag{7.48}$$

The representation as a block diagram of this system (Fig. 7.4) shows that the state x_3 is not observable.

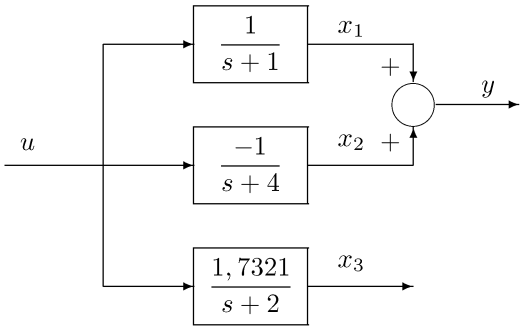


Figure 7.4: Unobservable system. Block diagram corresponding to the modal canonical form (Eq. (7.47))

7.4 Realizations

Any discrete system, described by the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, can also be represented by the discrete transfer function

$$H(z) = \boldsymbol{C}(z\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B} \tag{7.49}$$

The corresponding continuous-time model and transfer function could be considered as well.

It is only possible by a change of base to go from the controllability canonical form to the observability canonical form when the numerator and denominator of the system transfer function have no common root: the system is non degenerated. The set of the eigenvalues of $\boldsymbol{A}$ and the set of the poles of the transfer function are then identical. In this case, the system is controllable and observable. The state equation is then a minimal equation and constitutes a minimal

realization of the transfer function. Nonminimal realizations use a state vector of larger dimension than necessary, given the system transfer function. If a pole and a zero can cancel themselves, the concerned pole is either not controllable, or not observable.

In the most general case, a system can be decomposed into an observable and controllable part, an observable uncontrollable part, a controllable unobservable part and an unobservable and uncontrollable part. When unobservable or uncontrollable parts exist, the transfer function $H(z)$ can be simplified.

The study of controllability and of observability is, in fact, frequently delicate by use of controllability and observability matrices. The controllability and observability grammians possess better numerical properties and, moreover, their physical meaning is clearer. For a continuous-time system with constant coefficients, the controllability grammian[3] is defined by

$$G_{\mathcal{C}} = \int_0^\infty \exp(\boldsymbol{A}\tau)\boldsymbol{B}\boldsymbol{B}^T \exp(\boldsymbol{A}^T\tau)d\tau \tag{7.50}$$

and the observability grammian by

$$G_{\mathcal{O}} = \int_0^\infty \exp(\boldsymbol{A}\tau)\boldsymbol{C}\boldsymbol{C}^T \exp(\boldsymbol{A}^T\tau)d\tau \tag{7.51}$$

Practically, the system is studied for a finite time $[0, T]$ and the transient grammians are the integrals taken between the boundaries 0 and T, e.g. for the transient controllability grammian

$$G_{\mathcal{C}} = \int_0^T \exp(\boldsymbol{A}\tau)\boldsymbol{B}\boldsymbol{B}^T \exp(\boldsymbol{A}^T\tau)d\tau \tag{7.52}$$

which has as a limit the controllability grammian when $T \to \infty$. The controllability grammian (resp. observability) has maximum rank when the system is controllable (resp. observable). The number of uncontrollable states (resp. unobservable) corresponds to the number of singular values smaller than a given threshold ϵ.

The controllability grammian is the solution of the Lyapunov equation

$$\boldsymbol{A}G_{\mathcal{C}} + G_{\mathcal{C}}\boldsymbol{A}^T + \boldsymbol{B}\boldsymbol{B}^T = 0 \tag{7.53}$$

and the observability grammian is the solution of the Lyapunov equation

$$\boldsymbol{A}G_{\mathcal{O}} + G_{\mathcal{O}}\boldsymbol{A}^T + \boldsymbol{C}\boldsymbol{C}^T = 0 \tag{7.54}$$

For a discrete system with constant coefficients, the transient controllability grammian is defined by

$$G_{\mathcal{C},n} = \sum_{i=0}^{n-1} \boldsymbol{A}^i\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{A}^{Ti} \quad , \quad n \geq 1 \tag{7.55}$$

[3]The adjoint matrix $\boldsymbol{A}^*$is the transposed conjugate matrix of $\boldsymbol{A}$, equal to the transposed matrix $\boldsymbol{A}^T$ of $\boldsymbol{A}$ in the classical case of real-coefficient systems. For this reason, equations concerning the grammian are sometimes presented with the adjoint matrix instead of the transposed one, which has here been retained.

and the transient observability grammian by

$$G_{\mathcal{O},n} = \sum_{i=0}^{n-1} \boldsymbol{A}^i \boldsymbol{C}\boldsymbol{C}^T \boldsymbol{A}^{T\,i} \qquad , \quad n \geq 1 \tag{7.56}$$

Similarly to the continuous case, these transient grammians have as a limit the controllability or observability grammians when $n \to \infty$. The condition of controllability or of observability is that the corresponding grammian, which is a symmetrical positive matrix, is a defined matrix[4].

The physical meaning of the notion of grammian appears when the energy necessary to make a system go from a state $\boldsymbol{x} = 0$ to a state $\boldsymbol{x}_n$ is considered, which is equal to

$$E_n = \sum_{i=0}^{n-1} u_i^T u_i = \boldsymbol{x}_n^T (\boldsymbol{\mathcal{C}}_n \boldsymbol{\mathcal{C}}_n^T)^{-1} \boldsymbol{x}_n \tag{7.57}$$

with

$$\begin{array}{ll} \mathcal{C}_n & = \left[\boldsymbol{B} \quad \boldsymbol{A}\boldsymbol{B} \quad \ldots \quad \boldsymbol{A}^{n-1}\boldsymbol{B}\right] \\ G_{\mathcal{C},n} & = \boldsymbol{\mathcal{C}}_n \boldsymbol{\mathcal{C}}_n^T \end{array} \tag{7.58}$$

As the grammian $G_{\mathcal{C},n}$ can be defined by the recurrence

$$G_{\mathcal{C},n+1} = \boldsymbol{A}\, G_{\mathcal{C},n}\, \boldsymbol{A}^T + \boldsymbol{B}\, \boldsymbol{B}^T \tag{7.59}$$

the controllability grammian is the solution of the Lyapunov equation

$$G_{\mathcal{C}} = \boldsymbol{A}\, G_{\mathcal{C}}\, \boldsymbol{A}^T + \boldsymbol{B}\, \boldsymbol{B}^T \tag{7.60}$$

Similarly, the observability grammian is the solution of the Lyapunov equation

$$G_{\mathcal{O}} = \boldsymbol{A}\, G_{\mathcal{O}}\, \boldsymbol{A}^T + \boldsymbol{C}\, \boldsymbol{C}^T \tag{7.61}$$

The grammians allow us to calculate the balanced realization corresponding to a given system. A realization is said to be balanced when it possesses diagonal and equal controllability and observability grammians. It is then possible to proceed to a model reduction by comparing the orders of magnitude of the diagonal terms of the grammian and by neglecting those which are lower than a given threshold. The model reduction finds its application in particular in the solving of H_∞ synthesis problems (Maciejowski, 1989).

Example 7.3: **Balanced Realization and Model Reduction**

Consider the continuous-time system having three poles of relatively different order of magnitude, described by the following transfer function

$$G(s) = \frac{1}{(s+5)(s+1)(s+0.1)} \tag{7.62}$$

Its state-space representation is

$$\boldsymbol{A} = \begin{bmatrix} -6.1 & -5.6 & -0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad ; \quad \boldsymbol{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad ; \quad \boldsymbol{C} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \tag{7.63}$$

[4]The symmetrical matrix $\boldsymbol{A}$ is defined if $\boldsymbol{x}^T\, \boldsymbol{A}\, \boldsymbol{x} = 0 \Longrightarrow \boldsymbol{x} = 0$.

The controllability and observability matrices of this system are respectively equal to

$$\mathcal{C} = \begin{bmatrix} 1 & -6.1 & 31.61 \\ 0 & 1 & -6.1 \\ 0 & 0 & 1 \end{bmatrix} \quad ; \quad \mathcal{O} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{7.64}$$

and both have rank 3.

The controllability and observability grammians of system (7.63) are respectively equal to

$$\boldsymbol{G}_{\mathcal{C}} = \begin{bmatrix} 0.0832 & 0 & -0.0149 \\ 0 & 0.0149 & 0 \\ -0.0149 & 0 & 0.1812 \end{bmatrix} ; \boldsymbol{G}_{\mathcal{O}} = \begin{bmatrix} 0.1812 & 1.1055 & 1 \\ 1.1055 & 6.7582 & 6.1906 \\ 1 & 6.1906 & 6.1527 \end{bmatrix} \tag{7.65}$$

The balanced realization calculated for this system (A, B, C) gives the new system

$$\begin{aligned} \boldsymbol{A}_e &= \begin{bmatrix} -0.0692 & -0.1620 & -0.0527 \\ 0.1620 & -0.9213 & -0.6722 \\ -0.0527 & 0.6722 & -5.1095 \end{bmatrix} ; \boldsymbol{B}_e = \begin{bmatrix} 0.3887 \\ -0.4160 \\ 0.1483 \end{bmatrix} \\ \boldsymbol{C}_e &= \begin{bmatrix} 0.3887 & 0.4160 & 0.1483 \end{bmatrix} \end{aligned} \tag{7.66}$$

and the grammian corresponding to this balanced realization is the diagonal matrix, the elements of which are ordered by decreasing order of magnitude

$$\boldsymbol{G}_e = \begin{bmatrix} 1.0918 & 0 & 0 \\ 0 & 0.0939 & 0 \\ 0 & 0 & 0.0022 \end{bmatrix} \tag{7.67}$$

To proceed to a model reduction, the diagonal elements are compared to the first element. A first reduction is performed, choosing as a criterion to eliminate the components $g(i)$ lower than $g(1)/100$. Thus, the first reduced model is obtained

$$\boldsymbol{A}_{r1} = \begin{bmatrix} -0.0686 & -0.1690 \\ 0.1690 & -1.0097 \end{bmatrix}; \boldsymbol{B}_{r1} = \begin{bmatrix} 0.3871 \\ -0.4355 \end{bmatrix}; \boldsymbol{C}_{r1} = \begin{bmatrix} 0.3871 & 0.4355 \end{bmatrix} \tag{7.68}$$

and the corresponding transfer function

$$G_{r1}(s) = \frac{0.0043s^2 - 0.0351s + 0.1957}{s^2 + 1.0784s + 0.0979} \tag{7.69}$$

which presents the following poles: -0.9784 and -0.1, and zeros: $4.0834 \pm 5.3665j$. The fast pole of the original system has thus been cancelled and the two remaining poles are very close to the corresponding poles of the complete system. On the other hand, two zeros have been added.

A second reduction is performed by taking as a criterion to eliminate components $g(i)$ lower than $g(1)/10$. The second reduced model is obtained

$$\boldsymbol{A}_{r2} = \begin{bmatrix} -0.0969 \end{bmatrix} \quad ; \quad \boldsymbol{B}_{r2} = \begin{bmatrix} 0.46 \end{bmatrix} \quad ; \quad \boldsymbol{C}_2 = \begin{bmatrix} 0.46 \end{bmatrix} \tag{7.70}$$

and the corresponding transfer function

$$G_{r2}(s) = \frac{-0.1835s + 0.1938}{s + 0.0969} \tag{7.71}$$

which presents the following pole: -0.0969 and zero: 1.0561. The two faster poles of the origin system have been cancelled. One zero has been added.

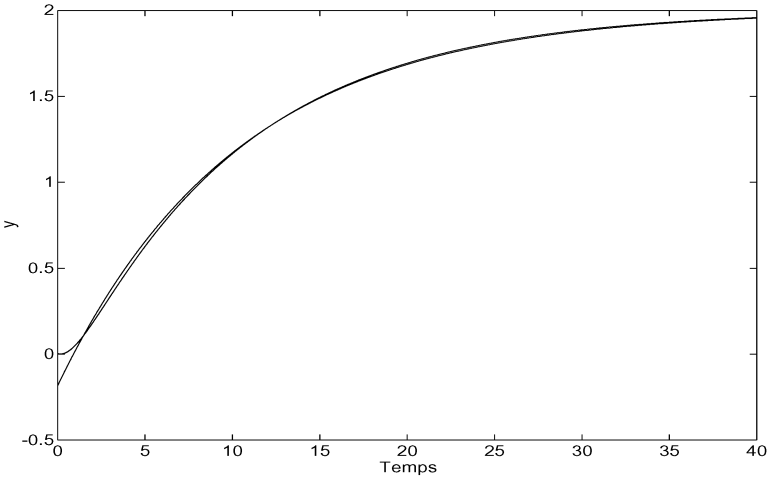


Figure 7.5: Model reduction: Response to a unit step, for the complete system and two reduced systems

A step response has been imposed on the original system, as well as on both reduced systems (Fig. 7.5). When only the fastest pole is cancelled, the response of this reduced second-order system is totally superposed to the response of the origin system (curve with inflection), while when two poles are cancelled, at very short times there appears a small deviation of the response (curve without inflection for the reduced first-order system), then the response becomes very close to the original one.

7.5 Remark on Controllability and Observability in Discrete Time

Assume that the continuous-time system is controllable and observable. A sufficient condition for the associated discrete-time system to be also controllable and observable is that the eigenvalues λ_i of the continuous state matrix $\boldsymbol{A}$ is such that:

- When $\mathcal{R}e(\lambda_i - \lambda_j) = 0$ with $i \neq j$
- Then $\mathcal{I}m(\lambda_i - \lambda_j) \neq \frac{2k\pi}{T_s}$, k being a nonzero relative integer, where T_s is the sampling period.

References

P. Borne, G. Dauphin-Tanguy, J.P. Richard, F. Rotella, and I. Zambetakis. *Modélisation et Identification des Processus*, volume 1, 2. Technip, Paris, 1992.

C. Foulard, S. Gentil, and J.P. Sandraz. *Commande et Régulation par Calculateur Numérique.* Eyrolles, Paris, 1987.

J.M. Maciejowski. *Multivariable Feedback Design.* Addison-Wesley, Wokingham, England, 1989.

W.M. Wohnam. *Linear Multivariable Control. A Geometric Approach.* Springer-Verlag, New York, 1985.

Part II

Multivariable Control

Chapter 8

Multivariable Control by Transfer Function Matrix

8.1 Introduction

The transfer function matrices are frequently used in chemical engineering, in particular because of the identification methods generally employed in this domain. In the present chapter, with the system being represented by a transfer function matrix, general characteristics relative to multi-input multi-output (MIMO) systems and to multivariable control are treated, such as the stability, interaction and decoupling.

Other chapters in the book using, in general, state-space modelling treat particular types of multivariable control: linear quadratic control and linear quadratic Gaussian control in Chap. 14, model predictive control in Chap. 16 and nonlinear multivariable control in Chap. 17. Notice that the state-space approach is more natural and appropriate for multivariable use than the approach by transfer functions. The description of a linear state-space system is essentially realized in Chap. 7 together with the notions of controllability and observability.

Multivariable control by PID controllers and multivariable internal model control making use of transfer functions are nevertheless treated in this chapter.

8.2 Representation of a Multivariable Process by Transfer Function Matrix

A multivariable process admits n_u inputs and n_y outputs. In general, the number of inputs should be larger than or equal to the number of outputs so that the process is controllable. Thus, we will assume $n_u \geq n_y$.

The system is supposed to have been identified in continuous time by transfer functions. In general, this identification is performed by sequentially impos-

ing signals such as steps on each input u_i $(i = 1, \ldots, n_u)$ and recording the corresponding vector of the responses y_{ij} $(j = 1, \ldots, n_y)$. From each input-output couple (u_i, y_{ij}), a transfer function is deduced by a least-squares procedure.

In open loop, the n_y outputs y_i are linked to the n_u inputs u_j and to the n_d disturbances d_k by the following set of n_y linear equations

$$\begin{aligned} Y_1 &= G_{u11}\,U_1 + \ldots + G_{u1n_u}\,U_{n_u} + G_{d11}\,D_1 + \ldots + G_{d1n_d}\,D_{n_d} \\ &\vdots \\ Y_{n_y} &= G_{un_y1}\,U_1 + \ldots + G_{un_yn_u}\,U_{n_u} + G_{dn_y1}\,D_1 + \ldots + G_{dn_yn_d}\,D_{n_d} \end{aligned} \tag{8.1}$$

which will be written in open loop under condensed matrix form as

$$\boldsymbol{Y} = \boldsymbol{G}_u\,\boldsymbol{U} + \boldsymbol{G_d}\,\boldsymbol{D} \tag{8.2}$$

where $\boldsymbol{y}$ is the output vector, $\boldsymbol{u}$ is the input vector and $\boldsymbol{d}$ is the disturbance vector (the modelled disturbances), $\boldsymbol{G}_u$ is the rectangular matrix $n_y \times n_u$, the elements of which are the input-output transfer functions, and $\boldsymbol{G_d}$ is the rectangular matrix $n_y \times n_d$, the elements of which are the disturbance-output transfer functions. The diagonal elements G_{uii} of matrix $\boldsymbol{G}_u$ represent the principal effects between the inputs and the outputs, while the non-diagonal elements represent the couplings.

From now on, it will be assumed, as in most frequent cases, that the number of inputs is equal to the number of outputs: the system is called square. In closed loop, it is also assumed that the control makes use of as many feedback controllers as inputs. Let $\boldsymbol{C}$ be the controller transfer function matrix. Industrially, it is common that a multivariable process is, in fact, only controlled by single-input single-output controllers; in this case, the matrix $\boldsymbol{C}$ contains only diagonal elements. It is also assumed that the actuators are represented by a transfer function matrix $\boldsymbol{G}_a$ and the measurement devices by a transfer function matrix $\boldsymbol{G}_m$.

The multivariable process can be represented by a block diagram, where each block symbolizes a transfer function matrix (Fig. 8.1).

The equations are the following

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{G}_u\,\boldsymbol{U} + \boldsymbol{G_d}\,\boldsymbol{D} \\ &= \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{E} + \boldsymbol{G_d}\,\boldsymbol{D} \\ &= \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,(\boldsymbol{K}_m\,\boldsymbol{Y_r} - \boldsymbol{G}_m\,\boldsymbol{Y}) + \boldsymbol{G_d}\,\boldsymbol{D} \end{aligned} \tag{8.3}$$

giving the closed-loop matrix relation

$$\boldsymbol{Y} = [\boldsymbol{I} + \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{G}_m]^{-1}\,\boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{K}_m\,\boldsymbol{Y_r} + [\boldsymbol{I} + \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{G}_m]^{-1}\,\boldsymbol{G_d}\,\boldsymbol{D} \tag{8.4}$$

In the previous expression, the matrix appearing by its inverse

$$\boldsymbol{F}_y(s) = \boldsymbol{I} + \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{G}_m \tag{8.5}$$

is called the feedback difference matrix measured at the output (Macfarlane and Belletrutti, 1973).

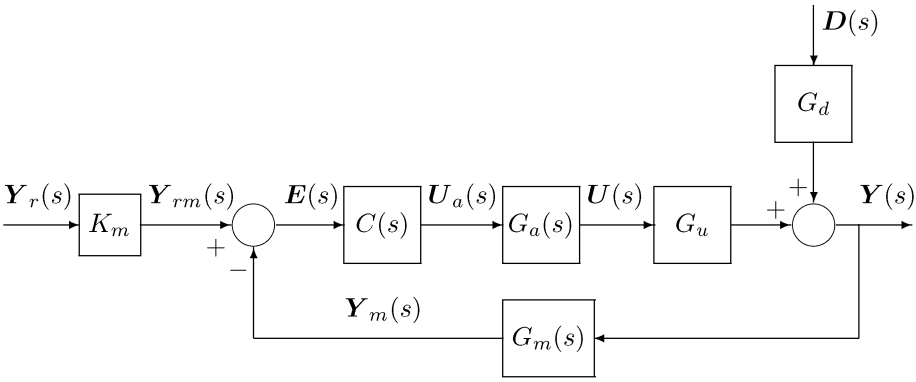


Figure 8.1: Block diagram of a multivariable process

The matrix

$$\boldsymbol{T}_y(s) = \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{G}_m \tag{8.6}$$

is called the feedback ratio matrix measured at the output.

The feedback differences and feedback ratios with respect to the input and to the error are also defined by

$$\begin{array}{ll} \boldsymbol{T}_u(s) & = \boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{G}_m\,\boldsymbol{G}_u \\ \boldsymbol{T}_e(s) & = \boldsymbol{G}_m\,\boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C} \\ \boldsymbol{F}_u(s) & = \boldsymbol{I} + \boldsymbol{T}_u(s) \\ \boldsymbol{F}_e(s) & = \boldsymbol{I} + \boldsymbol{T}_e(s) \end{array} \tag{8.7}$$

The characteristic equation is thus the expression of the cancellation of the feedback difference matrix determinant

$$\det[\boldsymbol{F}_y(s)] = \det[\boldsymbol{I} + \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{G}_m] = 0 \tag{8.8}$$

In fact, it can be verified that

$$\det[\boldsymbol{F}_y(s)] = \det[\boldsymbol{F}_u(s)] = \det[\boldsymbol{F}_e(s)] \tag{8.9}$$

The closed-loop matrix transfer function $\boldsymbol{G}_{bf}(s)$ relating $\boldsymbol{Y}_r(s)$ and $\boldsymbol{Y}(s)$ is equal to

$$\begin{array}{ll} \boldsymbol{G}_{bf}(s) & = \boldsymbol{F}_y^{-1}(s)\,\boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C} \\ & = \boldsymbol{G}_u\,\boldsymbol{F}_u^{-1}(s)\,\boldsymbol{G}_a\,\boldsymbol{C} \\ & = \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C}\,\boldsymbol{F}_e^{-1}(s) \end{array} \tag{8.10}$$

and the matrix

$$\boldsymbol{G}_{ol}(s) = \boldsymbol{G}_u\,\boldsymbol{G}_a\,\boldsymbol{C} \tag{8.11}$$

is called the open-loop transfer function matrix.

8.3 Stability Study

As in the single-input single-output case, the roots of the characteristic Eq. (8.8) must be located in the complex left half-plane so that the system is closed-loop stable. Several methods allow us to ensure this (Macfarlane and Belletrutti, 1973; Maciejowski, 1989). By definition, a transfer function matrix is exponentially stable if and only if it is proper and has no poles in the right half-plane.

These results can be transposed from continuous to discrete time by replacing the right half-plane by the inside of the unit circle.

8.3.1 Smith-McMillan Form

Let $\boldsymbol{G}(s)$ be a transfer function matrix, which is not necessarily square, of rank r. By a series of operations on the rows and the columns (Maciejowski, 1989), $\boldsymbol{G}(s)$ can be transformed into a pseudo-diagonal transfer function matrix $\boldsymbol{M}(s)$, such that

$$\boldsymbol{M}(s) = \text{diag}\{\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_r(s)}{\psi_r(s)}, 0, \ldots, 0\} \tag{8.12}$$

where the polynomials $\epsilon_i(s)$ and $\psi_i(s)$ are monic, coprime and satisfy the divisibility property

$$\begin{array}{ll} \epsilon_i(s) \text{ divides } \epsilon_{i+1}(s) & i = 1, \ldots, r-1 \\ \psi_{i+1}(s) \text{ divides } \psi_i(s) & i = 1, \ldots, r-1 \end{array} \tag{8.13}$$

$\boldsymbol{M}(s)$ is the Smith-McMillan form of $\boldsymbol{G}(s)$.

8.3.2 Poles and Zeros of a Transfer Function Matrix

Given the Smith-McMillan form of the transfer function matrix $\boldsymbol{G}(s)$, the roots of the polynomial product of the ψ_i

$$P(s) = \psi_1(s) \ \ldots \ \psi_r(s) \tag{8.14}$$

are the poles of $G(s)$ and the roots of the polynomial product of the ϵ_i

$$Z(s) = \epsilon_1(s) \ \ldots \ \epsilon_r(s) \tag{8.15}$$

are the zeros of $\boldsymbol{G}(s)$.

8.3.3 Generalized Nyquist Criterion

For a single-input single-output system, Cauchy's theorem and the resulting Nyquist criterion have been already commented on in Sect. 5.7.

For a multi-input multi-output system, square transfer function matrices will be considered (same numbers of inputs and outputs).

From the characteristic Eq. (8.8), the useful complex variable function for the Nyquist criterion is

$$F(s) = \det[\boldsymbol{I} + \boldsymbol{G}_u \boldsymbol{G}_a \boldsymbol{C} \boldsymbol{G}_m] \tag{8.16}$$

Suppose that this function possesses Z zeros and P poles in the right half-plane. As for single-input single-output systems, the algebraic number of encirclings (counted positively clockwise, negatively in the opposite sense) of the origin, when s describes the Nyquist contour, is equal to

$$N = Z - P \tag{8.17}$$

For the closed-loop system to be stable, it is necessary that: $Z = 0$.

Frequently, the open-loop process is stable, thus the transfer functions of matrix $\boldsymbol{G}_u$ have their poles in the left half-plane. It is assumed that the actuators and measurement devices are also stable. Moreover, the chosen controllers are also (preferably) stable. Assuming that the transfer functions are expressed in the form of rational fractions, the poles of $T(s) = \boldsymbol{G}_u \boldsymbol{G}_a \boldsymbol{C} \boldsymbol{G}_m$ are those of the function $F(s)$, so that if the process is open-loop stable, the function $F(s)$ has all its poles in the left half-plane. It follows that the number of zeros of this function in the right half-plane is then equal to the number of encirclings of the origin. In these conditions (stability of $\boldsymbol{G}_u$, $\boldsymbol{G}_a$, $\boldsymbol{C}$, $\boldsymbol{G}_m$), the important conclusion is that if the function $\det[\boldsymbol{I} + \boldsymbol{G}_u \boldsymbol{G}_a \boldsymbol{C} \boldsymbol{G}_m]$ encircles the origin, the process is closed-loop unstable.

In order to keep the analogy with the single-input single-output systems, in fact, the function is represented by

$$F(j\omega) = -1 + \det[\boldsymbol{I} + \boldsymbol{G}_u \boldsymbol{G}_a \boldsymbol{C} \boldsymbol{G}_m] \tag{8.18}$$

and the number of encirclings of the point $(-1, 0)$ is examined when the pulsation ω varies from $-\infty$ to $+\infty$.

8.3.4 Characteristic Loci

For a large-scale system, the Nyquist multivariable locus (Desoer and Wang, 1980; Macfarlane and Belletrutti, 1973), can be complicated and difficult to interpret. It is interesting to plot the eigenvalue loci, also called the characteristic loci.

Suppose that all the controllers are identical and equal to a steady-state gain k. The function $F(s)$ becomes equal to

$$F(s) = \det[\boldsymbol{I} + k \boldsymbol{G}_u \boldsymbol{G}_a \boldsymbol{G}_m] \tag{8.19}$$

The poles of F(s) are the poles of $\boldsymbol{G}_u \boldsymbol{G}_a \boldsymbol{G}_m$. The condition of stability is then that the trajectory of $F(s)$ encircles the origin P times anti-clockwise, when s describes the Nyquist contour.

It could be thought that it is necessary to draw the locus of $F(s)$ for each value of k. Indeed, if $\lambda_i(s)$ is an eigenvalue of $\boldsymbol{G}_u \boldsymbol{G}_a \boldsymbol{G}_m$, $k\lambda_i(s)$ is an eigenvalue

of $k\boldsymbol{G}_u\boldsymbol{G}_a\boldsymbol{G}_m$ and $1+k\lambda_i(s)$ is an eigenvalue of $\boldsymbol{I}+k\boldsymbol{G}_u\boldsymbol{G}_a\boldsymbol{G}_m$. As the function $F(s)$ is a determinant, it verifies

$$F(s) = \prod_i [1 + k\lambda_i(s)] \tag{8.20}$$

thus

$$\arg[F(s)] = \sum_i \arg[1 + k\lambda_i(s)] \tag{8.21}$$

The number of encirclings of the origin of $1 + k\lambda_i(s)$ could be counted; it is preferable to count the encirclings of the point $(-1, 0)$ of $k\lambda_i(s)$. The graphs are, in general, drawn for $k = 1$ when ω varies from $-\infty$ to $+\infty$ and the graphs of $\lambda_i(s)$ are called the characteristic loci. Individually, these graphs do not form a closed curve but, as a collective, they constitute a set of closed loops, as the imaginary eigenvalues are necessarily conjugate.

The generalized Nyquist criterion is thus formulated:
If the function $\boldsymbol{G}_u(s)\boldsymbol{G}_a(s)\boldsymbol{G}_m(s)$ possesses P unstable poles, the closed-loop system with the return difference $k\boldsymbol{G}_u(s)\boldsymbol{G}_a(s)\boldsymbol{G}_m(s)$ is stable if and only if the set of the characteristic loci encircles the point $(-1, 0)$ P times anti-clockwise (assuming that no hidden unstable modes exist).

To go from the characteristic loci obtained with $k = 1$ to those corresponding to any gain, it is sufficient to perform an homothety of centre 0 and ratio k. The limits of stability have been determined with $k = 1$ by examining the possible encirclings; the limit values of the gain k are easily deduced. The encirclings by $\boldsymbol{G}(s)$ of the point $(-1/k, 0)$ could also be considered.

The generalized inverse Nyquist criterion is obtained by reasoning with respect to ${\boldsymbol{G}_{\boldsymbol{ol}}}^{-1}(s) = [\boldsymbol{G}_u(s)\boldsymbol{G}_a(s)\boldsymbol{G}_m(s)]^{-1}$. If $\lambda_i(s)$ is an eigenvalue of $\boldsymbol{G}(s)$, $1/\lambda_i(s)$ is an eigenvalue of $\boldsymbol{G}^{-1}(s)$. The generalized inverse Nyquist criterion states that if $\boldsymbol{G}(s)$ has Z unstable transmission zeros, the closed loop system is stable if and only if the set of the characteristic loci of $\boldsymbol{G}^{-1}(s)/k$ encircles the point $(-1, 0)$ Z times anti-clockwise (assuming that no hidden unstable modes exist).

8.3.5 Gershgorin Circles

According to the Gershgorin theorem, given a complex matrix $\boldsymbol{A}$ of dimension $(n \times n)$, its eigenvalues λ are at least located in one of the discs defined by

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \tag{8.22}$$

and in one of the discs defined by

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}| \tag{8.23}$$

A Nyquist array of $\boldsymbol{G}_{ol}(s)$ is obtained by drawing an array of the Nyquist loci of all the elements $g_{ij}(s)$, where each locus referenced (i, j) occupies the position (i, j) in the array of the graphs.

The diagonal graphs, thus corresponding to the elements $g_{ii}(s)$, are considered. In each point of the graph of $g_{ii}(s)$, a circle can be placed with the following radius

$$\sum_{\substack{j=1 \\ j\neq i}}^{n_u} |g_{ij}(\mathrm{j}\omega)| \qquad \text{or} \qquad \sum_{\substack{j=1 \\ j\neq i}}^{n_u} |g_{ji}(\mathrm{j}\omega)| \tag{8.24}$$

The set of these circles forms a band enclosing the locus of $G_{ii}(\mathrm{j}\omega)$, called the Gershgorin band, and the circles are the Gershgorin circles.

If the Gershgorin bands of a matrix do not contain the origin, this matrix is diagonally dominant. The more diagonally dominant a matrix is, the narrower the Gershgorin bands are. To obtain the stability, it is necessary that $[\boldsymbol{I} + \boldsymbol{G}_{ol}(s)]$ is diagonally dominant. It is considered that $[\boldsymbol{I} + \boldsymbol{G}_{ol}(s)]$ is diagonally dominant if the Gershgorin bands of $\boldsymbol{G}_{ol}(s)$ do not encircle the point $(-1, 0)$. The more diagonally dominant $\boldsymbol{G}_{ol}(s)$ or $[\boldsymbol{I} + \boldsymbol{G}_{ol}(s)]$ is, the closer the multivariable system $\boldsymbol{G}_{ol}(s)$ is to a set of single-input single-output systems without interaction. In the case where no Gershgorin bands of $\boldsymbol{G}_{ol}(s)$ contain the point $(-1, 0)$, the closed-loop stability will be checked by counting the number of encirclings of the point $(-1, 0)$ by the bands, which is equivalent to using the characteristic loci for this particular case. However, one or several bands can contain the point $(-1, 0)$ without the system being unstable.

Similarly, the generalized inverse Nyquist criterion allows us to test the stability:
If all Gershgorin bands of $\boldsymbol{G}_{ol}^{-1}(s)$ exclude the point $(-1, 0)$, the stability will be checked by counting the number of encirclings of the point $(-1, 0)$ by the bands.

Rosenbrock (Maciejowski, 1989) gave a necessary and sufficient condition of stability:
Consider a system with n_u inputs and n_u outputs. Let $\boldsymbol{K} = \mathrm{diag}\{k_1, \ldots, k_{n_u}\}$ be the matrix of the controller gains. Suppose that the inequality

$$\left| g_{ii}(s) + \frac{1}{k_i} \right| > \sum_{j\neq i} |g_{ij}(s)| \qquad \forall i \tag{8.25}$$

is verified for all s of the Nyquist contour and that the i-th Gershgorin band of $\boldsymbol{G}(s)$ encircles the point $(-1/k_i, 0)$ N_i times anti-clockwise. The closed-loop system with return difference $-\boldsymbol{G}(s)\boldsymbol{K}$ is stable if and only if

$$\sum_i N_i = P \tag{8.26}$$

(assuming that no hidden unstable modes exist).

8.3.6 Niederlinski Index

This method can be applied only when PI controllers are used in the control loops. The method is especially used to avoid pairings of variables (an input linked to an output) which would be unsatisfactory. It also allows us to test the stability in some cases. Grosdidier et al. (1985) corrected the condition of Niederlinski. The Niederlinski index uses only steady-state gains of the transfer function matrix of the process $\boldsymbol{G}_u$. Suppose that $\boldsymbol{G}_u$, $\boldsymbol{G}_a$ and $\boldsymbol{G}_m$ are stable and note that $\boldsymbol{G} = \boldsymbol{G}_u\boldsymbol{G}_a$. The controller matrix $\boldsymbol{C}$ is diagonal and the gain of each controller is positive. The matrix $\boldsymbol{G}_{ol} = \boldsymbol{G}\boldsymbol{C}$ is rational and proper. All control systems taken individually are stable. The Niederlinski index is equal to

$$\frac{\det[\boldsymbol{G}(0)]}{\prod_i G_{ii}(0)} \tag{8.27}$$

where G_{ij} is the transfer function relating the input j and the output i. If this index is negative, the system will be unstable whatever the tuning of the controllers. If it is positive, it is impossible to conclude. Thus, it is a sufficient condition, except for multivariable systems of size lower than or equal to 2, where it is also necessary.

8.4 Interaction and Decoupling

A multivariable system presents the particularity that the inputs are coupled to the outputs. Different methods exist, allowing us to ensure at least a partial decoupling for a multivariable system. This is particularly important in the treatment by a transfer function matrix, but does not intervene as explicitly in linear quadratic Gaussian control (Chap. 14) or in model predictive control (Chap. 16). Most of the methods presented here are based on the fact that the interaction is undesired. But if the fact that a set point variation acts simultaneously on several outputs is annoying, this is not necessarily the case concerning the disturbance action on the outputs. In general, the most important point is that the process outputs follow the set point correctly, whatever the disturbances. A decoupling based on the set points in reality can make the disturbance rejection more difficult. The different methods presented must always be evaluated under these two aspects: set point tracking and disturbance rejection.

8.4.1 Decoupling for a 2×2 System

The aim of decoupling is to make the outputs y_i independent of the variation of any input u_j with $i \neq j$. Thus, for the 2×2 system, the output y_1 will be only influenced by the variations of u_1, thus of the set point y_{c1} and not of u_2. This result is obtained by adding, for example, to the controller a decoupling element De (Fig. 8.2).

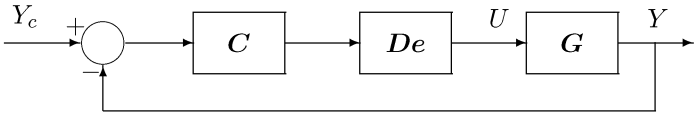


Figure 8.2: Interposition of a decoupler between the controller matrix and the process matrix

In the absence of decoupling, the open-loop equations are the following

$$\begin{array}{ll} Y_1 & = G_{u11}\,U_1 + G_{u12}\,U_2 + G_{d1}\,D_1 + \ldots \\ Y_2 & = G_{u21}\,U_1 + G_{u22}\,U_2 + G_{d2}\,D_2 + \ldots \end{array} \tag{8.28}$$

or in matrix form

$$\boldsymbol{Y} = \boldsymbol{G_u}\,\boldsymbol{U} + \boldsymbol{G_d}\,\boldsymbol{D} \tag{8.29}$$

and in closed loop

$$\boldsymbol{Y} = [\boldsymbol{I} + \boldsymbol{G_u}\,\boldsymbol{G_a}\,\boldsymbol{C}\,\boldsymbol{G_m}]^{-1}\,\boldsymbol{G_u}\,\boldsymbol{G_a}\,\boldsymbol{C}\,\boldsymbol{Y_r} + [\boldsymbol{I} + \boldsymbol{G_u}\,\boldsymbol{G_a}\,\boldsymbol{C}\,\boldsymbol{G_m}]^{-1}\,\boldsymbol{G_d}\,\boldsymbol{D} \tag{8.30}$$

In the presence of decouplers represented by a matrix $\boldsymbol{De}$ interposed between the controller matrix and the process transfer function matrix, this equation becomes

$$\begin{array}{rcl} \boldsymbol{Y} & = & [\boldsymbol{I} + \boldsymbol{G_u}\,\boldsymbol{G_a}\,\boldsymbol{De}\,\boldsymbol{C}\,\boldsymbol{G_m}]^{-1}\,\boldsymbol{G_u}\,\boldsymbol{G_a}\,\boldsymbol{De}\,\boldsymbol{C}\,\boldsymbol{Y_r} + \\ & & [\boldsymbol{I} + \boldsymbol{G_u}\,\boldsymbol{G_a}\,\boldsymbol{De}\,\boldsymbol{C}\,\boldsymbol{G_m}]^{-1}\,\boldsymbol{G_d}\,\boldsymbol{D} \\ & = & \boldsymbol{A_1}\,\boldsymbol{Y_r} + \boldsymbol{A_2}\,\boldsymbol{D} \end{array} \tag{8.31}$$

To avoid the interaction during set point variations, the matrix $\boldsymbol{A_1}$ must be diagonal. This type of decoupling that does not take into account disturbances is not strongly recommended. The following discussions will detail the reasons for the limitations of such decouplers.

Example 8.1: **Application to Wood and Berry Distillation Column**

Wood and Berry (1973) identified the transfer functions of a distillation column considered as a 2×2 system, having as inputs the reflux and reboiler vapour flow rates and as outputs the distillate and bottom mole fractions. The influence of the feed flow rate disturbance has been added to this model (Deshpande and Ash, 1988). The complete model is the following

$$\begin{bmatrix} Y_D(s) \\ Y_B(s) \end{bmatrix} = \begin{bmatrix} \dfrac{12.8\,\exp(-s)}{16.7s+1} & \dfrac{-18.9\,\exp(-3s)}{21s+1} \\ \dfrac{6.6\,\exp(-7s)}{10.9s+1} & \dfrac{-19.4\,\exp(-3s)}{14.4s+1} \end{bmatrix} \begin{bmatrix} R(s) \\ V(s) \end{bmatrix} + \begin{bmatrix} \dfrac{3.8\,\exp(-8s)}{14.9s+1} \\ \dfrac{4.9\,\exp(-3s)}{13.2s+1} \end{bmatrix} F(s) \tag{8.32}$$

The chosen steady-state regime is as follows:

$y_D = 0.96$, $y_B = 0.02$, $R = 1.95$ lb/min, $V = 1.71$ lb/min, $F = 2.45$ lb/min. This system was simulated in closed loop in the presence or in the absence of decoupling. The controllers are PI and the tunings are those cited by (Luyben, 1990), obtained by the method of largest modulus search (Sect. 8.7.1):
$K_{r1} = 0.375$, $\tau_{I1} = 8.29$, $K_{r2} = -0.075$, $\tau_{I2} = 23.6$

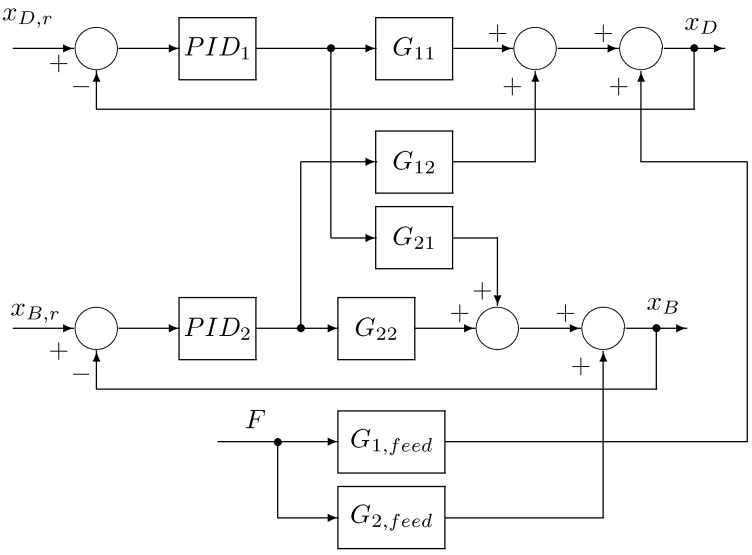


Figure 8.3: Wood and Berry distillation column. Closed-loop representation without decoupling

The closed-loop system without decoupling is represented in Fig. 8.3.

The system is first subjected to a distillate set point variation equal to -0.02 (Fig. 8.4). The bottom mole fraction strongly decreases before coming back to its nominal regime after a long period (about 120 min). Also note the influence of the time delays. The same system is subjected to a bottom product set point variation equal to 0.01 (Fig. 8.4). The distillate mole fraction is influenced, but more weakly than the bottom of the column in the previous case.

The same system is subjected to a feed flow rate disturbance step of 0.05 lb/min occurring after 20 min (Fig. 8.5). The bottom mole fraction is neatly more influenced than the top one.

The closed-loop system with decoupling is represented in Fig. 8.6. The decouplers have been calculated by the simplified decoupling. They are equal to

$$D_{12} = \frac{-G_{12}}{G_{11}} \quad ; \quad D_{21} = \frac{-G_{21}}{G_{22}} \tag{8.33}$$

As time delays exist, it is important that the decouplers are physically realizable. For example, hadd the time delay related to G_{11} been larger than the time delay related to G_{12}, the time delays would not have been taken into account. The closed-loop system with the same controllers but with decoupling is subjected to the same set point variations and the same disturbance. Figure 8.7

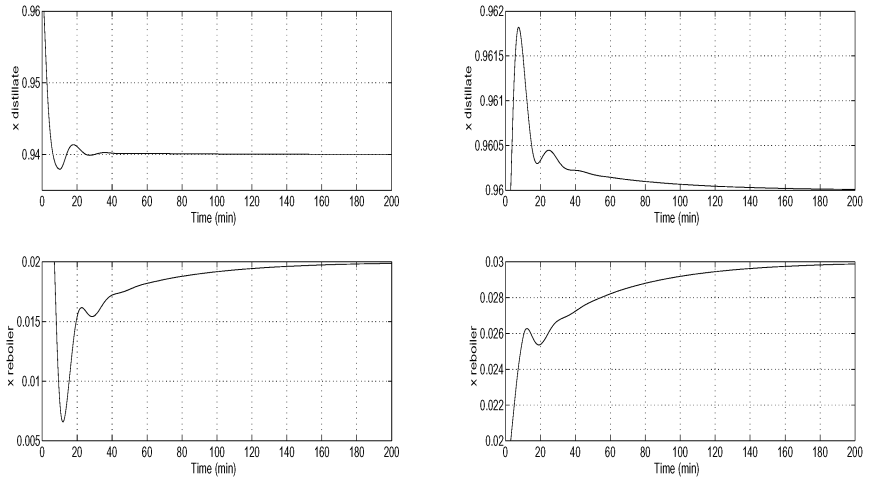


Figure 8.4: Wood and Berry distillation column without decoupling. Set point step of distillate −0.02 (left column) and set point step at the bottom 0.01 (right column). Distillate and bottom mole fraction responses

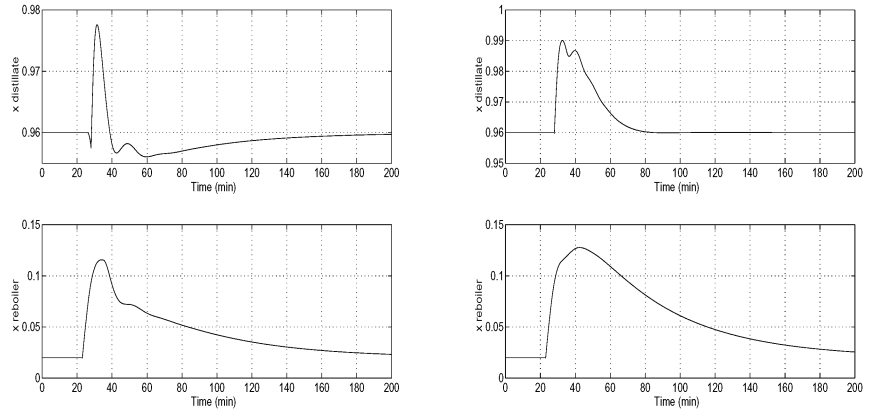


Figure 8.5: Wood and Berry distillation column without decoupling (left column) and with simplified decoupling (right column). Disturbance step of the feed flow rate 0.05 lb/min occurring after 20 min. Distillate and bottom product mole fraction responses

shows that the decoupling effectively acts by totally cancelling the multivariable effect. However, it must be noted that these results have been obtained in simulation and assume a perfect model. In reality, the transfer function

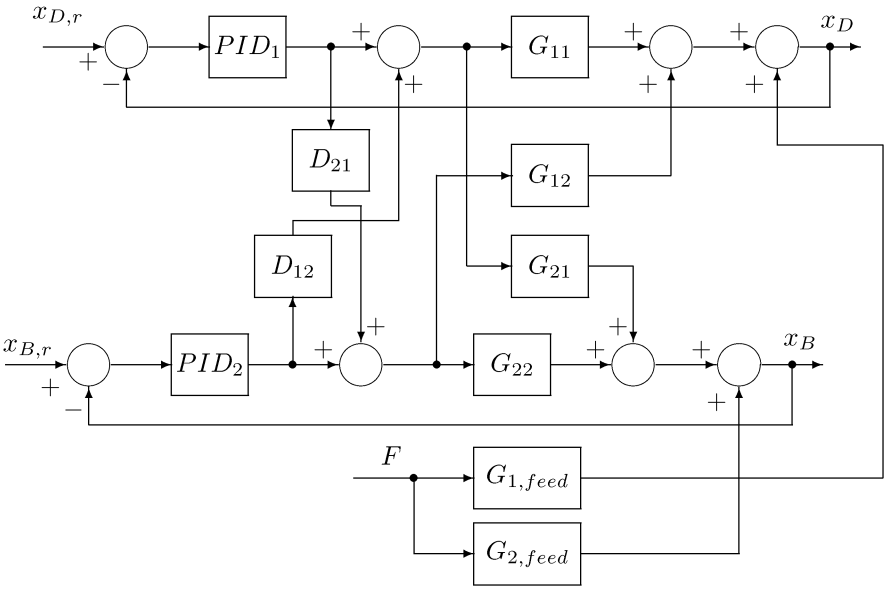


Figure 8.6: Wood and Berry distillation column. Closed-loop representation with decoupling

is known with uncertainty as well as the time delays. Thus, in practice, an influence of the coupling would be observed.

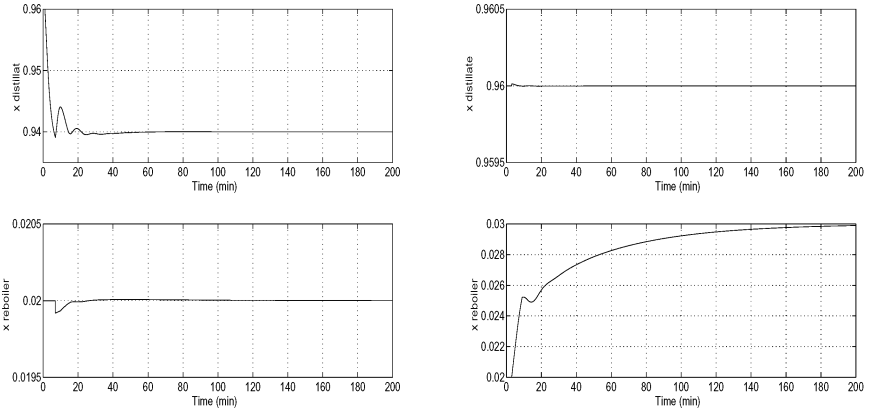


Figure 8.7: Wood and Berry distillation column with decoupling. Distillate set point step -0.02 (left column) and bottom set point step 0.01 (right column). Distillate and bottom product mole fraction responses

As the decouplers are calculated without taking into account the disturbances,

in Fig. 8.5 we note that the system with decoupling is no better with regard to disturbances than the system without decoupling; in the present case, it is even the contrary. Thus, this controller is not robust.

8.4.2 Disturbance Rejection

If a disturbance has no influence on a controlled variable, the modulus of the closed-loop transfer function is zero. The objective in the multivariable domain is to choose the best pairings between inputs and outputs so that the smallest moduli are obtained. Thus, for a multivariable system of dimension m and for each disturbance, the m curves must be plotted in the Bode diagram corresponding to the modulus of the m components of the vector $[\boldsymbol{I} + \boldsymbol{G_u}\,\boldsymbol{G_a}\,\boldsymbol{C}\,\boldsymbol{G_m}]^{-1}\,\boldsymbol{G_d}$.

8.4.3 Singular Value Decomposition

The decomposition of a matrix $\boldsymbol{A}$ of dimension $n \times p$ with respect to singular values is performed according to

$$\boldsymbol{A} = \boldsymbol{V}\,\boldsymbol{\Sigma}\,\boldsymbol{W}^T$$

where $\boldsymbol{V}$ is an orthonormal[1] matrix $n \times n$, $\boldsymbol{W}$ is an orthonormal matrix $p \times p$, and $\boldsymbol{\Sigma}$ is a diagonal matrix $n \times p$ with $\sigma_{ij} = 0$ if $i \neq j$ and $\sigma_{ii} = \sigma_i \geq 0$. The quantities σ_i are the singular values of $\boldsymbol{A}$ and the columns V_i of $\boldsymbol{V}$, respectively W_i of $\boldsymbol{W}$, are the left singular vectors, respectively right, such that

$$\boldsymbol{A}W_i = \sigma_i V_i \qquad \forall i \tag{8.34}$$

If $\boldsymbol{A}$ is a square matrix, the matrices $\boldsymbol{A}^T\,\boldsymbol{A}$ and $\boldsymbol{A}\,\boldsymbol{A}^T$ have the same nonzero eigenvalues.

The singular values of a rectangular matrix $\boldsymbol{A}$ of dimension $n \times p$ are the square roots of the eigenvalues of the square matrix $\boldsymbol{A}^*\,\boldsymbol{A}$ and thus are defined by

$$\sigma_i = \sqrt{\lambda_{i\,[\boldsymbol{A}^*\,\boldsymbol{A}]}}$$

Perform the singular value decomposition $\boldsymbol{G} = \boldsymbol{Y}\boldsymbol{\Sigma}\boldsymbol{U}^T$ and arrange the singular values in order of decreasing value: $\sigma_1 > \ldots > \sigma_{n_u}$. The columns of $\boldsymbol{U}$ denoted by $U_1, \ldots, U_{n_u}$ and the columns of $\boldsymbol{Y}$ denoted by $Y_1, \ldots, Y_{n_u}$

[1] A matrix $\boldsymbol{A}$ is orthonormal when it is:
1) Orthogonal: the set of vectors (columns or rows) composing the matrix is orthogonal, thus these vectors are orthogonal between themselves.
2) These vectors are unitary

$$\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{I}$$

where $\boldsymbol{A}^*$ represents the transposed conjugate matrix of a matrix $\boldsymbol{A}$.

respectively constitute the principal directions of the inputs and the outputs of $\boldsymbol{G}$. With the matrix $\boldsymbol{U}$ being orthonormal, the principal directions U_i are orthogonal between themselves; the same applies to the Y_j. If the input is chosen in the direction of U_i, so $u = \alpha U_i$ with the scalar α such that $|\alpha| = 1$, from the equality $y = \boldsymbol{G}u$, results

$$y = \alpha \sigma_i Y_i \tag{8.35}$$

the corresponding output is in direction Y_i and the gain of the system is σ_i. With each principal gain is thus associated a couple of principal directions. Note that, although this calculation can be performed for a non-square system, here this simplification is chosen.

The largest elements (in absolute value) in the columns of $\boldsymbol{Y}$ indicate which are the most sensitive inputs (Luyben, 1990). Deshpande and Ash (1988) mentions that the pairing should be realized by pairing the controlled output associated with the largest (in absolute value) element of Y_1 with the input associated with the largest (in absolute value) element of U_1, the same applies to the following directions U_2, Y_2, and so on.

8.4.4 Relative Gain Array

The interaction between loops can be evaluated by a method based on the study of the relative gain array (RGA) introduced by Bristol (1966). The loops influence themselves in a more or less important manner and a possible effect is that some loops destabilize the closed-loop system. The RGA method is relatively easy to implement, and for this reason is frequently used in chemical engineering. It is limited in its original form, as it uses only steady-state information. It can be extended by using frequency representations (Hovd and Skogestad, 1992; Skogestad and Morari, 1987b).

Steady-state Relative Gain Array

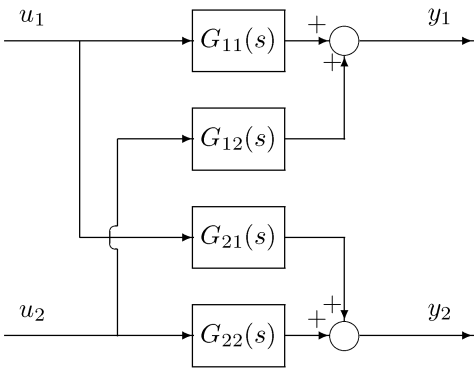


Figure 8.8: Block diagram of an open-loop multivariable 2×2 process

The simplest case that can be studied is that of a system with two inputs and two outputs considered around a steady state (Fig. 8.8)

$$\begin{aligned} Y_1(s) = &\quad G_{11}(s)\,U_1(s) + G_{12}(s)\,U_2(s) \\ Y_2(s) = &\quad G_{21}(s)\,U_1(s) + G_{22}(s)\,U_2(s) \end{aligned} \tag{8.36}$$

giving the open-loop transfer function between the input u_1 and the output y_1

$$\left(\frac{Y_1(s)}{U_1(s)}\right)_{ol} = G_{11}(s) \tag{8.37}$$

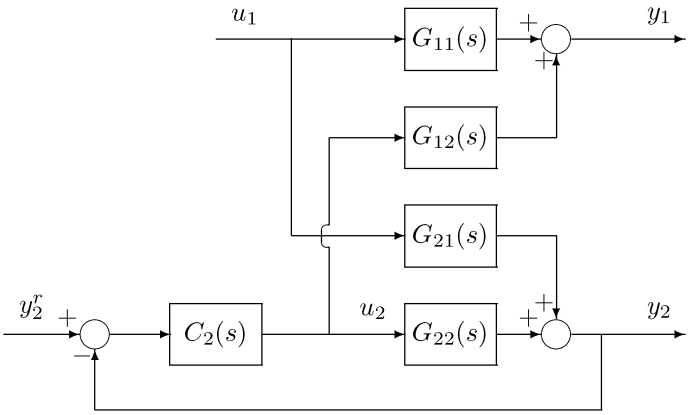


Figure 8.9: Block diagram of a multivariable 2×2 process with a controller on loop 2

Then suppose that only the output y_2 is controlled by installing a controller with transfer function $C_2(s)$ (Fig. 8.9). The system becomes equal to

$$\begin{aligned} Y_1(s) = &\quad \left[G_{11}(s) - \frac{G_{12}(s)C_2(s)G_{21}(s)}{1 + G_{22}(s)C_2(s)}\right] U_1(s) + \frac{G_{12}(s)C_2(s)}{1 + G_{22}(s)C_2(s)}\,Y_2^r(s) \\ Y_2(s) = &\quad \frac{G_{21}(s)}{1 + G_{22}(s)C_2(s)}\,U_1(s) + \frac{G_{22}(s)C_2(s)}{1 + G_{22}(s)C_2(s)}\,Y_2^r(s) \end{aligned} \tag{8.38}$$

The transfer function between u_1 and y_1 is thus modified (the subscript $cl22$ indicates that the loop between u_2 and y_2 is closed according to

$$\left(\frac{Y_1(s)}{U_1(s)}\right)_{cl22} = G_{11}(s) - \frac{G_{12}(s)C_2(s)G_{21}(s)}{1 + G_{22}(s)C_2(s)} \tag{8.39}$$

The ratio of the open-loop and closed-loop transfer functions expresses the influence of the loop between u_2 and y_2, so

$$\mu_{11}(s) = \frac{G_{11}(s)[1 + G_{22}(s)C_2(s)]}{G_{11}(s) + C_2(s)[G_{11}(s)G_{22}(s) - G_{12}(s)G_{21}(s)]} \tag{8.40}$$

Grosdidier et al. (1985) assume, as is frequently done, that the controller contains an integral action, which implies $C_2(0) = +\infty$, and the measure of the interaction $\mu_{11}(s)$ is denoted by λ_{11}, which depends only on the steady-state gains of the transfer function $\boldsymbol{G}$, thus

$$\lambda_{11} = \frac{G_{11}(0)\, G_{22}(0)}{G_{11}(0)\, G_{22}(0) - G_{12}(0)\, G_{21}(0)} \tag{8.41}$$

Explanation in terms of asymptotic variations:
The previous calculation can be found by reasoning for the two-input two-output system of Fig. 8.9. The asymptotic value of a variable is denoted by $x(\infty)$ to emphasize that time tends towards infinity. When the second loop is closed and the regulation is perfect, if a step is executed at u_1, it induces an action $G_{21}(0)\Delta u_1(\infty)$ at y_2 which must be compensated for by u_2 so that

$$G_{21}(0)\Delta u_1(\infty) + G_{22}(0)\Delta u_2(\infty) = 0 \qquad \Longrightarrow \qquad \Delta u_2(\infty) = -\frac{G_{21}(0)}{G_{22}(0)}\, \Delta u_1(\infty) \tag{8.42}$$

The action of the input variation $\Delta u_2(\infty)$ at y_1 is

$$G_{12}(0) \left[-\frac{G_{21}(0)}{G_{22}(0)}\, \Delta u_1(\infty) \right] \tag{8.43}$$

hence the total variation of y_1 due to the step at u_1

$$\begin{aligned} \Delta y_1(\infty) &= G_{11}(0)\, \Delta u_1(\infty) - \frac{G_{12}(0)\, G_{21}(0)}{G_{22}(0)}\, \Delta u_1(\infty) \\ &= \frac{G_{11}(0)\, G_{22}(0) - G_{12}(0)\, G_{21}(0)}{G_{22}(0)}\, \Delta u_1(\infty) \end{aligned} \tag{8.44}$$

The closed-loop ratio results

$$\left(\frac{\Delta y_1(\infty)}{\Delta u_1(\infty)} \right)_{cl} = \frac{G_{11}(0)\, G_{22}(0) - G_{12}(0)\, G_{21}(0)}{G_{22}(0)} \tag{8.45}$$

which, when coupled to the open-loop ratio

$$\left(\frac{\Delta y_1(\infty)}{\Delta u_1(\infty)} \right)_{ol} = G_{11}(0) \tag{8.46}$$

gives the first component of the steady-state relative gain array

$$\lambda_{11} = \frac{\left(\dfrac{\Delta y_1(\infty)}{\Delta u_1(\infty)} \right)_{ol}}{\left(\dfrac{\Delta y_1(\infty)}{\Delta u_1(\infty)} \right)_{cl}} = \frac{G_{11}(0)\, G_{22}(0)}{G_{11}(0)\, G_{22}(0) - G_{12}(0)\, G_{21}(0)} \tag{8.47}$$

The other elements of the steady-state relative gain array can be easily deduced by means of relation (8.50).

By generalizing for a larger dimension of the multivariable system, any element λ_{ij} of the relative gain array $\mathbf{\Lambda}$ is equal to the ratio of the steady-state gain between the i-th output variable y_i and the j-th input variable when the system is free of any control over the steady-state gain between the same variables when the system is feedback-controlled by the other inputs, assuming that in the steady state all the other outputs are maintained at their nominal value (the controllers are integral)

$$\lambda_{ij} = \frac{\left(\dfrac{\partial Y_i}{\partial U_j}\right)_{U_k=0,k\neq j}}{\left(\dfrac{\partial Y_i}{\partial U_j}\right)_{Y_l=0,l\neq i}} \tag{8.48}$$

assuming that the matrix $\boldsymbol{G}^{ij}$, obtained from the matrix $\boldsymbol{G}$ by suppressing the i-th row and the j-th column, is non-singular.

Denoting by $\bar{G}_{ij}$ the element (i, j) of matrix $\boldsymbol{G}^{-1}$, the elements λ_{ij} of the relative gain array can be calculated (Shinskey, 1988) by the following relation

$$\lambda_{ij} = G_{ij}\,\bar{G}_{ji} \tag{8.49}$$

The elements λ_{ij} verify the relation

$$\sum_{i=1}^{n_u} \lambda_{ij} = \sum_{j=1}^{n_u} \lambda_{ij} = 1 \tag{8.50}$$

For a perfectly decoupled system, the relative gain array $\mathbf{\Lambda}$ becomes the identity matrix. Theoretically, the objective is that the associations (principal effects) between an output variable and an input variable are such that the elements of the concerned matrix are positive and are close to 1 (Bristol, 1966). When an element (in absolute value) is large, it means that the interaction is important.

For a MIMO system, the general recommendations (Seborg et al., 1989; Shinskey, 1988) are:

- If $\lambda_{ij} < 0$, the gain of the concerned pairing changes its sign when passing from open to closed loop. y_j should not be paired with u_i. For a two-input two-output system, if $\lambda_{11} < 0$, y_1 should not be paired with u_1.
- If $\lambda_{ij} \approx 0$, the interaction is low, y_i is liitemttle influenced by u_j. For a two-input two-output system, if $\lambda_{11} \approx 0$, y_2 should be paired with u_1.
- If $0 < \lambda_{ij} < 1$, the closed-loop gain $u_j - y_i$ is larger than the open-loop gain and the interaction between the loops is maximum for $\lambda_{ij} = 0.5$. For a two-input two-output system, y_1 should be paired with u_1 only if $\lambda_{11} > 0.5$.
- If $\lambda_{ij} > 1$, the interaction is high and the degree of interaction between the loops increases with λ_{ij}.

Integral Stabilizability

The use of the relative gain array assumes that the controllers comprise an integral action and can be decomposed into a matrix of integrators $k/s\boldsymbol{I}$ and a

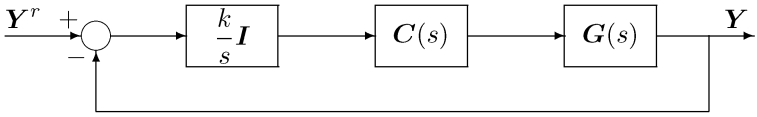


Figure 8.10: Integral multivariable control

matrix of compensators $\boldsymbol{C}(s)$ (Fig. 8.10). Assuming that the matrix $\boldsymbol{H}(s) = \boldsymbol{G}(s)\boldsymbol{C}(s)$ is a proper transfer function matrix ($\lim_{s\to\infty} \boldsymbol{G}(s) = 0$), the system defined by $\boldsymbol{H}(s)$ is stabilizable (Grosdidier et al., 1985) by an integral action only if

$$\det(\boldsymbol{H}(0)) > 0 \tag{8.51}$$

Integral Controllability
The system defined by $\boldsymbol{H}(s)$ is controllable by an integral action only if all the eigenvalues of $\boldsymbol{H}(0)$ are in the right complex half-plane. If only one of the eigenvalues is located in the left half-plane, the system is not controllable.
Relation with RGA matrix
When $\lambda_{jj}(\boldsymbol{G}) < 0$, then for any compensator $\boldsymbol{C}(s)$ such that:
a) $\boldsymbol{G}(s)\boldsymbol{C}(s)$ is proper,
b) y_j influences only u_j and u_j is influenced only by y_j, resulting in $C_{jl} = C_{lj} = 0$,
either the closed-loop system is unstable,
or the loop j is unstable, all the other loops being open,
or the closed-loop system is unstable when j is taken off.
Integrity
Grosdidier et al. (1985) also consider the necessary conditions so that a multivariable system remains stable when one of the actuators or sensors breaks down (integrity concept) and the coupled output is no more controlled: the system is tolerant with respect to a fault of the sensor or the actuator j if the complete and the reduced systems obtained by removing this sensor j are simultaneously controllable by an integral action.

Grosdidier et al. (1985), in particular, study the cases of 2×2 and 3×3 systems for which many incomplete results had been previously obtained. Thus, for a 2×2 system with $\lambda_{11} > 0$, there exists a diagonal compensator $\boldsymbol{C}(s)$ such that the closed-loop system remains stable in spite of the breakdown of a sensor or an actuator. For a 3×3 system with $\lambda_{jj} > 0, \forall j$, if there exists a compensator $\boldsymbol{C}(s)$ such that $\boldsymbol{H}(s)$ is controllable by an integral action and $H_{jj}(0) = 0$, then it is possible to maintain the stability of the system in spite of the breakdown. For a 2×2 system, a multivariable compensator is not necessary, while it should be used for a 3×3 system. It is recommended to use the concept of integral controllability previously exposed.
Robustness and relative gain array
The designer wishes that the control system remains stable even when the model $\tilde{\boldsymbol{G}}(s)$ that was used to design the multivariable compensator $\boldsymbol{C}(s)$ does not represent anymore perfectly the process $\boldsymbol{G}(s)$. Grosdidier et al. (1985)

show that the process running with the same compensator $\boldsymbol{C}(s)$ keeps integral controllability, provided that

$$\frac{\|\boldsymbol{G}(s)-\tilde{\boldsymbol{G}}(s)\|}{\|\tilde{\boldsymbol{G}}(s)\|} < \frac{1}{\|\tilde{\boldsymbol{G}}(s)\| \ \|\tilde{\boldsymbol{G}}^{-1}(s)\|} = \frac{1}{\mathrm{Cond}(\tilde{\boldsymbol{G}})} \tag{8.52}$$

where Cond is the condition number[2] of the matrix $\tilde{\boldsymbol{G}}$.

A problem is that the condition number depends on the scaling of the inputs and the outputs (realized by multiplication of $\boldsymbol{G}$ by diagonal matrices D_u and D_y) opposite to the RGA matrix. Nevertheless, it is possible (Grosdidier et al., 1985; Skogestad and Morari, 1987a), to realize a scaling so that a minimum or optimal condition number γ^* is obtained. For a 2×2 system, this minimum condition number can be related to the 1-norm of the RGA matrix by

$$\gamma^* = \|\Lambda_1\| + \sqrt{\|\Lambda_1\|^2 - 1} \tag{8.53}$$

and the following conjecture is proposed (Grosdidier et al., 1985) for a square multivariable system of any order

$$\gamma^* \leq 2 \max[\|\Lambda_1\|, \|\Lambda_\infty\|] \tag{8.54}$$

which is verified for a second-order system. This rejoins the results concerning the stability of 2×2 systems for which Shinskey (1979) showed that if $\lambda_{11} < 0$ or $\lambda_{11} > 1$, the error of the decoupler necessary to destabilize the system decreases when $|\lambda_{11}|$ increases although this error has no influence on the stability when $0 < \lambda_{11} < 1$.

Recalling that $\bar{G}_{ij}$ are the elements of matrix $\boldsymbol{G}^{-1}$, the relative variations of G_{ij}, $\bar{G}_{ij}$, λ_{ij} are related by

$$\frac{d\lambda_{ij}}{\lambda_{ij}} = (1-\lambda_{ij})\,\frac{dG_{ij}}{G_{ij}} \qquad , \qquad \frac{d\lambda_{ij}}{\lambda_{ij}} = \frac{\lambda_{ij}-1}{\lambda_{ij}}\,\frac{d\bar{G}_{ij}}{\bar{G}_{ij}} \tag{8.55}$$

From Eq. (8.55), when some elements of the RGA matrix Λ are large compared to 1, it results that the system is sensitive to modelling errors. If the

[2] Numerically, often the ratio of the largest singular value to the smallest singular value is considered as a measure of the condition number denoted by $\gamma(\boldsymbol{A})$ of the matrix. When the orders of magnitude of the singular values differ strongly, this means that the problem is ill-conditioned, or still that the equations of the associated linear system present some dependence. For a linear system $\boldsymbol{AX} = \boldsymbol{B}$, the solution $\boldsymbol{X}$ would vary largely for small variations of $\boldsymbol{B}$.

A matrix norm subordinate to a vector norm is defined by

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|}$$

Thus, the 1-norm denoted by $\|\boldsymbol{A}\|_1$ is the maximum with respect to the columns of the sum of the absolute values of the elements of a column, and the ∞-norm denoted by $\|\boldsymbol{A}\|_\infty$ is the maximum with respect to the rows of the sum of the absolute values of the elements of a row. The subordinate matrix norm (sometimes without subscript in this chapter) is the Euclidean norm or 2-norm denoted by $\|\boldsymbol{A}\|_2$ equal to the largest singular value of $\boldsymbol{A}$, where the Euclidean norm of vector $\boldsymbol{x}$ is equal to $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^*\boldsymbol{x}}$. The largest singular value is equal to the 2-norm of the matrix.

norm of Λ is large, and from Eq. (8.54) the minimum condition number is large, then the system is practically uncontrollable (Skogestad and Morari, 1987b). A system is sensitive to uncertainties on the control variable, whose variations are never perfectly known, when the process and the controller simultaneously have large elements in the RGA matrix (Skogestad and Morari, 1987b). This could explain the sensitivity of decouplers, in particular those called ideal or simplified, when the process has large elements in the RGA matrix. In the same way, the use of a controller based on the inverse of the model must be avoided in these conditions. A one-way decoupler (triangular decoupling matrix) is far less sensitive to the uncertainty of the control variable and a diagonal decoupler is not at all sensitive, but the latter may provide very bad performances.

For 2×2 systems, the RGA matrix Λ and the Niederlinski index give the same information (Hovd and Skogestad, 1994).

Pairing rule for integrity

It is recommended, for stable processes, to choose pairings corresponding to positive values of the Niederlinski index and of the RGA matrix $\Lambda(0)$. Hovd and Skogestad (1994) give indications for processes presenting one or several unstable poles.

Dynamic Relative Gain Array

It is possible to represent the modulus of the elements of the relative gain array Λ with respect to frequency (Tung and Edgar, 1981). In the following robustness analysis, a systematic frequency study is realized.

For example, a controller having a relative gain array whose modulus of the elements is small at all frequencies is in general insensitive to uncertainties of the control variable.

8.4.5 Gershgorin Circles and Interaction

If the system were perfectly decoupled, the radii of the Gershgorin circles would be zero. The larger the radius, the wider the band and the greater the interaction. By successive approximations of the controllers, it is possible to reduce the bandwidth and to decouple the system at the maximum. In spite of all this, the problem of disturbance rejection is not present in this method hence its limited interest.

8.5 Multivariable Robustness

In single-input single-output systems, the modulus of the closed-loop transfer function is examined

$$\left| \frac{Y(j\omega)}{Y_r(j\omega)} \right| = \frac{G_{ol}(j\omega)}{1 + G_{ol}(j\omega)} \tag{8.56}$$

This flat curve at low frequencies (for which the ratio $|Y(j\omega)/Y_r(j\omega)|$ is equal to 1) presents a resonance peak for a given frequency ω_p, then decreases when the frequency increases. The more important the maximum, the less underdamped

the system and thus the less robust the system. More generally, the sensitivity and complementary sensitivity functions have been introduced and robustness criteria have been defined (Sect. 5.10) in order to account for uncertainties either in the system or in the process environment.

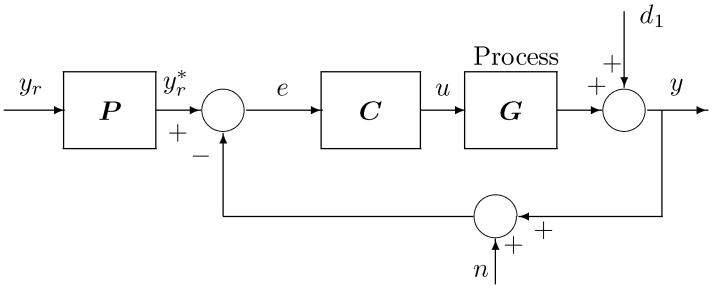


Figure 8.11: Block diagram of a multivariable process for robustness study

For a multivariable system, robustness has been developed in a similar way (Doyle and Stein, 1981; Maciejowski, 1989; Oustaloup, 1994; Palazoglu and Arkun, 1985; Skogestad and Postlethwaite, 1996). In particular, we will examine the necessary compromise between performance, i.e. the capability of the system to follow a reference trajectory, and robustness, whose objective is to react correctly in the face of uncertainties. The uncertainties, being generally weak at low frequencies, increase largely with frequency. Different types of uncertainties can be considered, e.g. additive uncertainty and multiplicative uncertainty equivalent to a relative variation. Taking into account an uncertainty means that the perturbed system keeps some properties, e.g. stability or performance, in spite of the undergone variation.

Consider the system in Fig. 8.11. The output is equal to

$$\begin{aligned} Y &= (\boldsymbol{I} + \boldsymbol{G}\,\boldsymbol{C})^{-1}\,\boldsymbol{G}\,\boldsymbol{C}\,(Y_r^* - N) + (\boldsymbol{I} + \boldsymbol{G}\,\boldsymbol{C})^{-1}\,D \\ &= \boldsymbol{T}(s)\,Y_r^* + \boldsymbol{S}(s)\,D - \boldsymbol{T}(s)\,N \\ &= (\boldsymbol{I} - \boldsymbol{S}(s))\,Y_r^* + \boldsymbol{S}(s)\,D - (\boldsymbol{I} - \boldsymbol{S}(s))\,N \\ &= \boldsymbol{T}(s)\,Y_r^* + (\boldsymbol{I} - \boldsymbol{T}(s))\,D - \boldsymbol{T}(s)\,N \end{aligned} \tag{8.57}$$

where the sensitivity function of the output is equal to

$$\boldsymbol{S}(s) = [\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1} \tag{8.58}$$

and the complementary sensitivity function

$$\begin{aligned} \boldsymbol{T}(s) &= [\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1}\,\boldsymbol{G}(s)\,\boldsymbol{C}(s) \\ &= \boldsymbol{S}(s)\,\boldsymbol{G}(s)\,\boldsymbol{C}(s) \\ &= \boldsymbol{I} - \boldsymbol{S}(s) \end{aligned} \tag{8.59}$$

The tracking error is equal to

$$\begin{aligned} E &= Y_r^* - Y \\ &= (\boldsymbol{I} + \boldsymbol{G}\,\boldsymbol{C})^{-1}\,(Y_r^* - D) + (\boldsymbol{I} + \boldsymbol{G}\,\boldsymbol{C})^{-1}\,\boldsymbol{G}\,\boldsymbol{C}\,N \\ &= \boldsymbol{S}(s)\,Y_r^* - \boldsymbol{S}(s)\,D + \boldsymbol{T}(s)\,N \end{aligned} \tag{8.60}$$

Denote by $\sigma_m(G)$ the smallest singular value of a matrix $\boldsymbol{G}$ and by $\sigma^M(G)$ the largest singular value, which verifies

$$\sigma^M(G) = \|\boldsymbol{G}\| \tag{8.61}$$

When the singular values are expressed with respect to frequency: $\sigma_i(\omega)$, they are called the principal gains of $\boldsymbol{G}(s)$.

Using the singular value decomposition, it can be shown that

$$\sigma_m(\omega) < \frac{\|\boldsymbol{G}(j\omega)U(j\omega)\|}{\|U(j\omega)\|} < \sigma^M(\omega) \tag{8.62}$$

thus the gain of a multivariable system is a value between the smallest and the largest principal gain.

Doyle and Stein (1981) use limits such as $l_m(\omega)$ to represent a boundary for uncertainty, low at low frequencies and high at high frequencies. Here, a simplified formulation is qualitatively retained by low or high, which in fact utilises the same concepts.

According to Eq. (8.57), the objective is to maintain the sensitivity $\boldsymbol{S}$ at a low value at low frequencies (for disturbance rejection), and what is important is the upper boundary represented by $\sigma^M(\boldsymbol{S})$

$$\sigma^M([\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1}) \quad \text{low} \tag{8.63}$$

On the other hand, the complementary sensitivity function $\boldsymbol{T}$ must be of a low value at high frequencies (to minimize the measurement noise influence) and the upper boundary represented by $\sigma^M(\boldsymbol{T})$ is of interest

$$\sigma^M(\boldsymbol{I} - [\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1}) \quad \text{low} \tag{8.64}$$

The compromise between performance and robustness is the inequality

$$|1-\sigma^M(\boldsymbol{S})| \le \sigma^M(\boldsymbol{T}) \le 1+\sigma^M(\boldsymbol{S}) \quad \text{and:} \quad |1-\sigma^M(\boldsymbol{T})| \le \sigma^M(\boldsymbol{S}) \le 1+\sigma^M(\boldsymbol{T}) \tag{8.65}$$

Define the output bandwidth ω_{by} for which

$$\sigma^M(\boldsymbol{T})(\omega_{by}) = 1/\sqrt{2}\sigma^M(\boldsymbol{T})(0) = 1/\sqrt{2}$$

In the case without the precompensator $\boldsymbol{P}$, the system has only one degree of freedom and if the transmission bandwidth ω_{ty} is designed such that

$$\sigma_m(\boldsymbol{T})(\omega_{ty}) = 1/\sqrt{2}\sigma_m(\boldsymbol{T})(0) = 1/\sqrt{2}$$

then it results that

$$\omega_{ty} < \omega_{by} \tag{8.66}$$

and the designer must seek to minimize $(\omega_{by} - \omega_{ty})$ for the set point tracking, which amounts to seeking

$$\sigma_m(\boldsymbol{I} - [\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1}) \approx 1 \quad \text{and} \quad \sigma^M(\boldsymbol{I} - [\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1}) \approx 1 \tag{8.67}$$

In the case with the precompensator $\boldsymbol{P}$, the system has two degrees of freedom; the transmission bandwidth is then determined by the matrix $\boldsymbol{TP}$, giving $\sigma_m(\boldsymbol{TP})$ and the designer must try to maximize $(\omega_{ty} - \omega_{by})$ for the set point tracking, as it is then possible to make $\omega_{ty} > \omega_{by}$.

The input is equal to

$$U = [\boldsymbol{I} + \boldsymbol{CG}]^{-1} \boldsymbol{C} \, (Y_r^* - N - D) \tag{8.68}$$

Denote the control sensitivity by

$$\boldsymbol{F}_i^{-1}(s) = [\boldsymbol{I} + \boldsymbol{C}(s)\boldsymbol{G}(s)]^{-1} \tag{8.69}$$

The variations of the input signal should be maintained as weak thus, $\sigma^M(\boldsymbol{F}_i^{-1}\boldsymbol{C})$ must be weak. As

$$\sigma^M(\boldsymbol{F}_i^{-1}\boldsymbol{C}) \leq \frac{\sigma^M(\boldsymbol{C})}{\sigma_m(\boldsymbol{F}_i)} \tag{8.70}$$

it suffices to make $\sigma^M(\boldsymbol{F}_i^{-1}\boldsymbol{C})$ small so that the ratio is small, which is only possible if $\sigma^M(\boldsymbol{G}) \gg 1$ or if $\sigma_m(\boldsymbol{C}) \ll 1$. At frequencies where $\sigma^M(\boldsymbol{G})$ is not large compared to 1, $\sigma_m(\boldsymbol{C})$ must be maintained as low as possible

$$\sigma_m(\boldsymbol{C}) \quad \text{low} \tag{8.71}$$

In general, it is interesting to express the previous criteria (8.63), (8.64), (8.67), (8.71) with respect to the principal gains of the open-loop transfer function matrix $\boldsymbol{G}_{ol} = \boldsymbol{GC}$ (Doyle and Stein, 1981; Maciejowski, 1989).

Disturbance rejection:

The criterion (8.63) dealing with low sensitivity at low frequencies can be expressed again as

$$\sigma^M([\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1}) \leq \frac{1}{\sigma_m(\boldsymbol{G}(s)\,\boldsymbol{C}(s))} \tag{8.72}$$

thus a low sensitivity is obtained when the smallest principal gain of the open loop is large: $\sigma_m(\boldsymbol{GC}) \gg 1$.

Minimization of the measurement noise:

The criterion (8.64) dealing with the complementary sensitivity low at high frequencies can be expressed again as

$$\sigma^M([\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1}) = \frac{1}{\sigma_m(\boldsymbol{I} + [\boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1})} \tag{8.73}$$

thus a low sensitivity to measurement noise is obtained when the largest principal gain of the open loop is small: $\sigma^M(\boldsymbol{GC}) \ll 1$.

Set point tracking:

The criterion (8.67) obtained in the case of a system with only one degree of freedom can be expressed again as

$$\boldsymbol{I} - [\boldsymbol{I} + \boldsymbol{G}(s)\,\boldsymbol{C}(s)]^{-1} \approx 1 \tag{8.74}$$

which can be transformed into $\sigma_m(\boldsymbol{GC}) \gg 1$.
Small input variations:
The criterion (8.71) gave $\sigma^M(\boldsymbol{C}) \ll 1$.

The apparent conflicts between these four demands can be solved by considering the frequency domains where each of them must be applied.

The closed-loop performance can be deduced from the moduli of the characteristic loci when the matrix $\boldsymbol{GC}$ is normal[3]. If the matrix $\boldsymbol{GC}$ is not quite normal, the characteristic loci can still be used. If the deviation with respect to normality which can be calculated according to Maciejowski (1989) is large, the only indication is

$$\sigma_m \leq |\lambda_i| \sigma^M \tag{8.75}$$

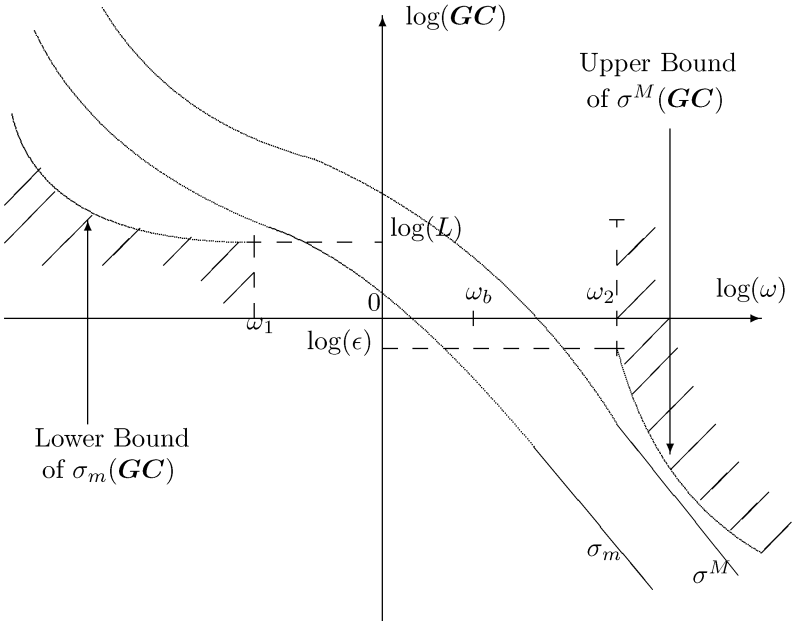


Figure 8.12: Specification representation according to the open-loop transfer function matrix $\boldsymbol{GC}$ of a multivariable process

The loci of the eigenvalues of $\boldsymbol{GC}$ can be drawn in the classical Bode representation and interpreted in the usual manner in order to deduce considerations of gain and phase

$$\begin{aligned}
&|\det(\boldsymbol{GC})| \geq L^{n_u} && \text{for: } \omega \leq \omega_1 && \text{with: } \sigma_m(\boldsymbol{GC}) \geq L \\
&|\det(\boldsymbol{GC})| \leq \epsilon^{n_u} && \text{for: } \omega \geq \omega_2 && \text{with: } \sigma^M(\boldsymbol{GC}) \leq \epsilon \\
&\arg\lambda_i(\boldsymbol{GC}) \leq -\frac{\pi}{2}\,\frac{\log(L/\epsilon)}{\log(\omega_2/\omega_1)} && \text{if: } \boldsymbol{GC} \text{ open-loop stable}
\end{aligned} \tag{8.76}$$

[3] A matrix $\boldsymbol{A}$ is normal when it verifies $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^*$.

This last equation allows us to obtain the following limit

$$\frac{20\log(L/\epsilon)}{\log(\omega_2/\omega_1)} < 40\text{dB/decade} \tag{8.77}$$

which, if it is not verified, leads to difficulties in maintaining the closed-loop stability.

The singular values $\sigma_m(\boldsymbol{GC})$ and $\sigma^M(\boldsymbol{GC})$ previously defined can also be drawn with respect to frequency in a similar way to a Bode diagram; these will be the loci of the singular values. The specifications (Fig. 8.12) dealing with the smallest and largest principal gains and the frequency bands can be applied to build the specifications (Doyle and Stein, 1981; Maciejowski, 1989) of $\boldsymbol{GC}$ similarly to those defined during the robustness study for single-input single-output systems (Kwakernaak, 1993). The frequency constraints provided by the criteria to respect are then approximated by continuous transfer functions of simple expression which produce frequency weightings W_1 and W_2 such that the original specifications (8.63) and (8.64) become

$$\|W_1\boldsymbol{S}\|_\infty \ll 1 \quad \text{and:} \quad \|W_2\boldsymbol{T}\|_\infty \ll 1 \tag{8.78}$$

8.6 Robustness Study of a 2×2 Distillation Column

Example 8.2: **Robustness Analysis of a Distillation Column**

As an applied example of robustness analysis of a 2×2 system, consider the study of a distillation column by Arkun et al. (1984). The comments are taken from this paper. Refer also to the example developed in Sect. 20.4.2.

Denote by $\boldsymbol{De}$ the decoupling matrix of the system having transfer function matrix $\boldsymbol{G}$ (Fig. 8.2). The designer of the decoupler wishes that the following equation to be satisfied

$$\boldsymbol{G}(s)\,\boldsymbol{De}(s) = \boldsymbol{M}(s) \tag{8.79}$$

where $\boldsymbol{M}(s)$ is a diagonal matrix specified by the designer.

8.6.1 Simplified Decoupling Analysis

Consider the matrix $\boldsymbol{M}$

$$\boldsymbol{M} = \begin{bmatrix} G_{11}\left(1 - \dfrac{G_{12}\,G_{21}}{G_{11}\,G_{22}}\right) & 0 \\ 0 & G_{22}\left(1 - \dfrac{G_{12}\,G_{21}}{G_{11}\,G_{22}}\right) \end{bmatrix} \tag{8.80}$$

the matrix $\boldsymbol{G}$ being

$$\boldsymbol{G} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \tag{8.81}$$

A simplified decoupling is given by the decoupling matrix

$$\boldsymbol{De} = \begin{bmatrix} 1 & -\dfrac{G_{12}}{G_{11}} \\ -\dfrac{G_{21}}{G_{22}} & 1 \end{bmatrix} \tag{8.82}$$

which results in the following output equations

$$\begin{aligned} Y_1 &= G_{11}\left(1 - \frac{G_{12}\,G_{21}}{G_{11}\,G_{22}}\right) U_1 \\ Y_2 &= G_{22}\left(1 - \frac{G_{12}\,G_{21}}{G_{11}\,G_{22}}\right) U_2 \end{aligned} \tag{8.83}$$

As the minimum singular value $\sigma_m[\boldsymbol{I} + \boldsymbol{G}_{ol}(j\omega)] > 0$, the system is stable. Nevertheless, if both minimum and maximum singular values are lower than 1, the system is not robust with respect to sensitivity. Thus, high-purity distillation columns are more sensitive to modelling errors and the stability margin is smaller than that for low-purity columns (Arkun et al., 1984).

8.6.2 Ideal Decoupling Analysis

Consider the matrix $\boldsymbol{M}$

$$\boldsymbol{M} = \begin{bmatrix} G_{11} & 0 \\ 0 & G_{22} \end{bmatrix} \tag{8.84}$$

The ideal decoupling is given by the decoupling matrix

$$\boldsymbol{De} = \begin{bmatrix} \dfrac{1}{1 - \dfrac{G_{12}\,G_{21}}{G_{11}\,G_{22}}} & -\dfrac{\dfrac{G_{12}}{G_{11}}}{1 - \dfrac{G_{12}\,G_{21}}{G_{11}\,G_{22}}} \\ -\dfrac{\dfrac{G_{21}}{G_{22}}}{1 - \dfrac{G_{12}\,G_{21}}{G_{11}\,G_{22}}} & \dfrac{1}{1 - \dfrac{G_{12}\,G_{21}}{G_{11}\,G_{22}}} \end{bmatrix} \tag{8.85}$$

which results in the output equations

$$\begin{aligned} Y_1 &= G_{11}\,U_1 \\ Y_2 &= G_{22}\,U_2 \end{aligned} \tag{8.86}$$

By examining the singular values (Arkun et al., 1984), it can be shown that the simplified decoupling works better than the ideal decoupling.

8.6.3 One-Way Decoupling Analysis

Consider the matrix $\boldsymbol{M}$

$$\boldsymbol{M} = \begin{bmatrix} G_{11}\left(1 - \dfrac{G_{12}\,G_{21}}{G_{11}\,G_{22}}\right) & G_{12} \\ 0 & G_{22} \end{bmatrix} \tag{8.87}$$

The one-way decoupling is given by the matrix decoupling

$$\boldsymbol{De} = \begin{bmatrix} 1 & 0 \\ -\dfrac{G_{21}}{G_{22}} & 1 \end{bmatrix} \tag{8.88}$$

which results in the output equations

$$\begin{aligned} Y_1 &= G_{11}\left[1 - \frac{G_{12}\,G_{21}}{G_{11}\,G_{22}}\right] U_1 + G_{12}\,U_2 \\ Y_2 &= G_{22}\,U_2 \end{aligned} \tag{8.89}$$

8.6.4 Comparison of the Three Previous Decouplings

Note that

$$l_m(\omega) = \sigma_m[\boldsymbol{I} + [\boldsymbol{G}(j\omega)\,\boldsymbol{De}(j\omega)\,\boldsymbol{C}(j\omega)\,]^{-1}] \tag{8.90}$$

is the maximum tolerable degree of uncertainty. As l_m is maximum for the simplified decoupling, this simplified decoupling can tolerate more uncertainty than the other types of decoupling which were mentioned.

For a low-purity distillation column, all the decoupling schemes improve the robustness. For a high-purity column, the ideal decoupling makes the column less robust.

8.7 Synthesis of a Multivariable Controller

In all cases, first a choice of the manipulated variables and of the controlled variables must be realized. The reflection on the variable pairing can be partly based on the Niederlinski index and on the analysis of the relative gain array (RGA) (Hovd and Skogestad, 1994). It is possible to eliminate some impossible variable pairings by use of the Niederlinski index. The best pairings among those possible must be chosen by applying methods such as the RGA. Then, the synthesis of a multivariable control can be performed at very different complexity levels.

Thus, the multivariable aspect may be ignored by closing the loops individually at their turn: first a loop is closed while allowing $m - 1$ open loops, then a second loop while allowing $m - 2$ open loops, and so on. The multivariable controller is then diagonal. The method of the largest modulus proposed in Sect. 8.7.1 can be considered to be derived from this procedure.

Except for this case, if a real multivariable controller is designed, the method of the characteristic loci allows us to approach the control system.

The robustness analysis based on the study of the singular values should complete the synthesis in all cases, even if the study has been voluntarily simplified. The pairing could be chosen as the one which gives the smallest modulus for the closed-loop transfer function relative to the disturbance.

8.7.1 Controller Tuning by the Largest Modulus Method

The following method (Luyben, 1990, 1986) of search of the largest modulus is relatively simple to realize, presents the advantage of being easily understood and is parent to the Ziegler-Nichols method, which was used for single-input single-output controllers. It is decomposed into four stages:

- Calculation of the parameters of the PI controller according to the Ziegler-Nichols method for each individual loop (ultimate frequency ω_u and ultimate gain of each diagonal transfer function G_{ii}; at the ultimate frequency, the phase angle is $-\pi$ and the ultimate gain is then the inverse of the real part of G_{ii}. The gain of the controller recommended by Luyben (1990) is $K_{ci} = K_{ui}/2.2$ and the integral time constant is $\tau_{ci} = 2\pi/1.2\omega_{ui}$.
- A detuning factor $F > 1$ (from about 1.5 to 4) is chosen. All the controller gains are divided by F: $K_{ci} = K_{ZNi}/F$ (a safety margin is ensured). The controller time constants are multiplied by the same factor F: $\tau_{ci} = \tau_{ZNi} * F$ (also a safety margin by slowing the response).
  Represent the function

$$W(j\omega) = -1 + \mathrm{Det}\,[\boldsymbol{I} + \boldsymbol{G}_{ol}] \tag{8.91}$$

  in the complex plane. The closer the function is to the Nyquist point $(-1, 0)$, the closer it is to instability. By analogy with closed-loop SISO systems, the closed-loop multivariable modulus (here expressed in decibels) is defined

$$L = 20 \log |\frac{W}{1+W}| \tag{8.92}$$

  which presents a maximum with respect to frequency.
- The detuning factor is varied until the maximum of L: $L_{\max}$ is equal to $2n$, n being the order of the multivariable system: for a SISO system, this corresponds to the usual recommendation $L_{\max} = 2\mathrm{dB}$.

This method of the largest modulus guarantees not only the stability of the control system in its environment, but also that of each controller considered individually. Following this method of the largest modulus, it is quite possible to pursue the improvement of the control system.

8.7.2 Controller Tuning by the Characteristic Loci Method

Maciejowski (1989) proposes the method of the characteristic loci in agreement with Grosdidier et al. (1985) (Fig. 8.10), which can be decomposed into successive stages

- Calculate a constant compensator $\boldsymbol{C}_h \approx -\boldsymbol{G}^{-1}(\omega_b)$ (corresponding to high frequencies) where ω_b is the loop passband,
- Calculate an approximately commutative compensator $\boldsymbol{C}_m(s)$ at a medium frequency $\omega_m < \omega_b$ for the compensated process $\boldsymbol{G}(s)\boldsymbol{C}_h$ such that $\boldsymbol{C}_m(j\omega) \to \boldsymbol{I}$ when $\omega \to \infty$ (to avoid the influence of the high-frequency noise on the decoupling realized by $\boldsymbol{C}_m(s)$). In this stage, the characteristic loci are considered as well as the behaviour in the neighbourhood of the critical point $(-1, 0)$.

- In order to compensate for the possible steady-state errors, design an approximately commutative compensator $\boldsymbol{C}_i(s)$ at a low frequency $\omega_1 < \omega_m$ for the compensated process $\boldsymbol{G}(s)\boldsymbol{C}_h\boldsymbol{C}_m(s)$ such that $\boldsymbol{C}_i(j\omega) \to \boldsymbol{I}$ when $\omega \to \infty$. The integral action integral can be introduced in $\boldsymbol{C}_i(s)$.
- The complete compensator is equal to

$$C(s) = \boldsymbol{C}_h\boldsymbol{C}_m(s)\boldsymbol{C}_i(s) \tag{8.93}$$

8.8 Discrete Multivariable Internal Model Control

Among the different model-based controls can be found model algorithmic control, dynamic matrix control and internal model control. However, Garcia and Morari (1985) showed that the first two types of control can be classified under internal model control (Fig. 8.13).

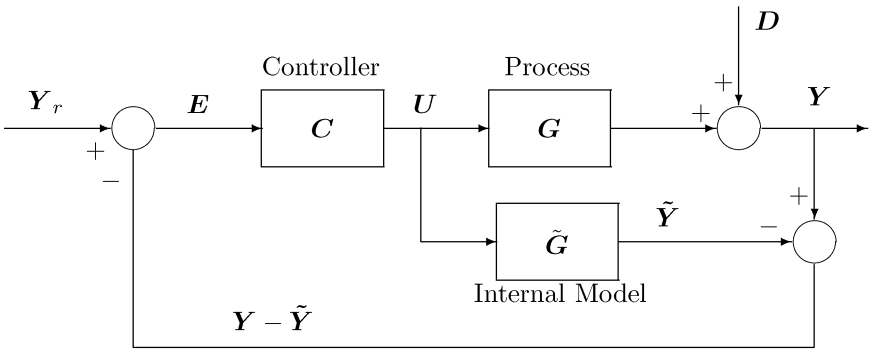


Figure 8.13: Discrete multivariable internal model control

Internal model control is developed for multivariable systems (Garcia and Morari, 1985) in the same way as for single-input single-output systems. Multivariable internal model control is here presented with reference to the z-transform. If continuous transfer function matrices in s were used, the reasoning would be similar, provided that the respective conditions of stability were kept. The input vector $\boldsymbol{u}$ of the system is equal to

$$\boldsymbol{U} = [\boldsymbol{I} + \boldsymbol{C}(z)\,(\boldsymbol{G}(z) - \tilde{\boldsymbol{G}}(z))]^{-1}\,\boldsymbol{C}(z)\,(\boldsymbol{Y}_r(z) - \boldsymbol{D}(z)) \tag{8.94}$$

and the output vector

$$\boldsymbol{Y} = \boldsymbol{G}(z)\,[\boldsymbol{I} + \boldsymbol{C}(z)\,(\boldsymbol{G}(z) - \tilde{\boldsymbol{G}}(z))]^{-1}\;\boldsymbol{C}(z)\,(\boldsymbol{Y}_r(z) - \boldsymbol{D}(z)) + \boldsymbol{D}(z) \tag{8.95}$$

If the model $\tilde{\boldsymbol{G}}$ and the process $\boldsymbol{G}(z)$ coincide perfectly, these equations are reduced to

$$\boldsymbol{U} = \boldsymbol{C}(z)\,(\boldsymbol{Y}_r(z) - \boldsymbol{D}(z)) \tag{8.96}$$

and

$$\boldsymbol{Y} = \boldsymbol{G}(z)\,\boldsymbol{C}(z)(\boldsymbol{Y}_r(z) - \boldsymbol{D}(z)) + \boldsymbol{D}(z) \tag{8.97}$$

The ideal controller corresponding to a zero steady-state error and a perfect disturbance rejection is obtained by taking

$$\boldsymbol{C}(z) = [\boldsymbol{G}(z)]^{-1} \qquad \Longrightarrow \quad \boldsymbol{C}(z) = [\tilde{\boldsymbol{G}}(z)]^{-1} \tag{8.98}$$

In fact, such a controller is not physically realizable if the model $\tilde{\boldsymbol{G}}(z)$ contains delays or transmission zeros outside the unit circle. Moreover, even if this is not the case, if the transmission zeros are too close to unity (by a lower value), the controller will not be robust and the output will present oscillations.

To avoid these problems, a factorization of the model transfer function is realized as

$$\tilde{\boldsymbol{G}}(z) = \tilde{\boldsymbol{G}}_+(z)\,\tilde{\boldsymbol{G}}_-(z) \tag{8.99}$$

so that $\tilde{\boldsymbol{G}}_+(z)$ contains the delays and the transmission zeros. Thus, $[\tilde{\boldsymbol{G}}_-(z)]^{-1}$ is a physically realizable controller and the controller is chosen as

$$\boldsymbol{C}(z) = [\tilde{\boldsymbol{G}}_-(z)]^{-1} \tag{8.100}$$

Delay factorization:
If all the delays t_d were identical, the problem would be simple and would result in

$$\tilde{\boldsymbol{G}}_+(z) = z^{-(t_d+1)}\,\boldsymbol{I} \tag{8.101}$$

where 1 unit is added to the delay to account for the fact that an output always responds to an input variation with a delay of at least one sampling period.

In general, the delays are different and denoted by t_{di} so that for n outputs it results that

$$\tilde{\boldsymbol{G}}_+(z) = \mathrm{diag}\,[z^{-(t^+_{d1}+1)} \ \ldots \ z^{-(t^+_{dn}+1)}] \tag{8.102}$$

by setting

$$t^+_{dj} = \max_i[\max\{0,\, t_{dij}\}] \tag{8.103}$$

The delays t_{dij} are found from the inverse of the model matrix $\tilde{\boldsymbol{G}}(z)^{-1}$ whose elements are $z^{t_{dij}+1}\,g_{ij}(z)$ such that $g_{ij}(z)$ is semi-proper (a ratio of polynomials $g_{ij}(z)$ is said to be semi-proper when the order of the numerator is lower or equal to that of the denominator).

In these conditions, in the absence of modelling errors, the output is equal to

$$\begin{aligned} \boldsymbol{y} &= \tilde{\boldsymbol{G}}(z)\,\tilde{\boldsymbol{G}}_-(z)^{-1}\,(\boldsymbol{y}_r(z) - \boldsymbol{d}(z)) + \boldsymbol{d}(z) \\ &= \tilde{\boldsymbol{G}}_+(z)\,(\boldsymbol{y}_r(z) - \boldsymbol{d}(z)) + \boldsymbol{d}(z) \end{aligned} \tag{8.104}$$

If the matrix $\tilde{\boldsymbol{G}}_+(z)$ is diagonal, the closed-loop outputs are decoupled.
Factorization of the transmission zeros out of the unit circle:
The controller $\tilde{\boldsymbol{G}}_-(z)$ must be stable. With the zeros of the model being poles for the controller, it is clear that these zeros must be located out of the unit circle.

$\tilde{\boldsymbol{G}}_-(z)$ is decomposed into a product of two factors and denoted by

$$\tilde{\boldsymbol{G}}_-(z) = \tilde{\boldsymbol{G}}_{-1}(z)\,\tilde{\boldsymbol{G}}_{+1}(z) \tag{8.105}$$

where $\tilde{\boldsymbol{G}}_{-1}(z)$ contains the zeros inside the unit circle that will be kept for the controller, and $\tilde{\boldsymbol{G}}_{+1}(z)$ contains the zeros outside the unit circle that are added to $\tilde{\boldsymbol{G}}_+(z)$.

To respect the wish that the matrix $\tilde{\boldsymbol{G}}_{+1}(z)$ is diagonal, the integral of the squared error criterion is minimized by the scalar diagonal matrix

$$\tilde{\boldsymbol{G}}_{+1}(z) = \prod_{i=1}^{m} \left(\frac{z-\nu_i}{z-\hat{\nu}_i}\right) \left(\frac{1-\hat{\nu}_i}{1-\nu_i}\right) \boldsymbol{I} \tag{8.106}$$

where ν_i are the transmission zeros and $\hat{\nu}_i$ are their images inside the unit circle such that

$$\hat{\nu}_i = \begin{cases} \nu_i & \text{if } |\nu_i| \leq 1 \\ \frac{1}{\nu_i} & \text{if } |\nu_i| > 1 \end{cases} \tag{8.107}$$

In fact, the ideal decoupling by diagonalization is not optimal for the integral of the squared error criterion and other non-diagonal factorizations can be preferable.

References

Y. Arkun, B. Manousiouthakis, and A. Palazoglu. Robustness analysis of process control systems. A case study of decoupling control in distillation. *Ind. Eng. Chem. Process Des. Dev.*, 23:93–101, 1984.

E.H. Bristol. On a new measure of interactions for multivariable process control. *IEEE Trans. Automat. Control*, AC-11:133–134, 1966.

P.B. Deshpande and R.A. Ash. *Computer Process Control with Advanced Control Applications*. Instrument Society of America, North Carolina, 2nd edition, 1988.

C.A. Desoer and Y.T. Wang. On the generalized Nyquist stability criterion. *IEEE Trans. Automat. Control*, AC-25(2):187–196, 1980.

J.C. Doyle and G. Stein. Multivariable feedback design: Concepts for a classical/modern synthesis. *IEEE Trans. Automat. Control*, AC-26:4–16, 1981.

C.E. Garcia and M. Morari. Internal model control. 2. Design procedure for multivariable systems. *Ind. Eng. Chem. Process Des. Dev.*, 24:472–484, 1985.

P. Grosdidier, M. Morari, and B.R. Holt. Closed-loop properties from steady-state gain information. *Ind. Eng. Chem. Fundam.*, 24:221–235, 1985.

M. Hovd and S. Skogestad. Simple frequency-dependent tools for control system analysis, structure selection and design. *Automatica*, 28(5):989–996, 1992.

M. Hovd and S. Skogestad. Pairing criteria for decentralized control of unstable plants. *Ind. Eng. Chem. Res.*, 33:2134–2139, 1994.

H. Kwakernaak. Robust control and $\mathcal{H}\infty$-optimization – tutorial paper. *Automatica*, 29(2):255–273, 1993.

W. L. Luyben. *Process Modeling, Simulation, and Control for Chemical Engineers*. McGraw-Hill, New York, 1990.

W.L. Luyben. Simple method for tuning SISO controllers in multivariable systems. *Ind. Eng. Chem. Process Des. Dev.*, 25:654–660, 1986.

A.G.J. Macfarlane and J.J. Belletrutti. The characteristic locus design method. *Automatica*, 9:575–588, 1973.

J.M. Maciejowski. *Multivariable Feedback Design*. Addison-Wesley, Wokingham, England, 1989.

A. Oustaloup, editor. *La Robustesse. Analyse et Synthèse de Commandes Robustes*. Hermès, Paris, 1994.

A. Palazoglu and Y. Arkun. Robust tuning of process control systems using singular values and their sensitivities. *Chem. Eng. Commun.*, 37:315–331, 1985.

D.E. Seborg, T.F. Edgar, and D.A. Mellichamp. *Process Dynamics and Control*. Wiley, New York, 1989.

F.G. Shinskey. *Process Control Systems*. McGraw-Hill, New York, 1979.

F.G. Shinskey. *Process Control Systems*. McGraw-Hill, New York, 3rd edition, 1988.

S. Skogestad and M. Morari. Control configuration selection for distillation columns. *AIChE J.*, 33(10):1620–1635, 1987a.

S. Skogestad and M. Morari. Implications of large RGA elements on control performance. *Ind. Eng. Chem. Res.*, 26:2323–2330, 1987b.

S. Skogestad and I. Postlethwaite. *Multivariable Feedback Control. Analysis and Design*. Wiley, Chichester, 1996.

L.S. Tung and T.F. Edgar. Analysis of control-output interactions in dynamic systems. *AIChE J.*, 27(4):690–693, 1981.

R.K. Wood and M.W. Berry. Terminal composition control of a binary distillation column. *Chem. Eng. Sci.*, 28:1707–1717, 1973.

Part III

Discrete-Time Identification

Chapter 9

Discrete-Time Generalities and Basic Signal Processing

The use of computer or, more generally, discrete time implies large differences in the way of approaching the process control problem compared to continuous time. As a matter of fact, a non-negligible part of the study will have to be devoted to the realization of a measurement interface, allowing us to express the analog signals delivered by the sensors in a digital form adapted to calculation (analog to digital conversion: A/D), as well as to express in analog form the digital inputs delivered by the computer (digital to analog conversion: D/A). The computer always contains a supervising system, which is essential, and allows the operator to control or intervene in the process.

When time is continuous and when the signal is continuous with regard to its amplitude, the signal is analog, such as the raw signal delivered by a physical sensor. Due to the data acquisition system, measurements are realized at regular time intervals. Time is no more treated as a continuous variable, but is discretized; for this reason, the corresponding time-dependent signals constitute variables. A discrete-time signal with a continuous amplitude is said to be sampled. Moreover, a consequence of use of computer is that the variable obtained by this means cannot take any value, as it is digitalized; between adjacent amplitudes there exists a finite nonzero interval related to the quantization (8 bits, 12 bits or 16 bits, for example), and this is called quantization of the variable. A digital signal is constituted by a quantized variable in discrete time.

Digital signals need the use of the z-transform, which plays the role of the Laplace transform for continuous signals.

9.1 Fourier Transformation and Signal Processing

The important use of Fourier transformation in signal processing justifies a short review at this stage.

9.1.1 Continuous Fourier Transform

The Fourier transform of a continuous function $f(t)$ is defined by

$$\mathcal{F}(f(t)) = \hat{f}(\nu) = \int_{-\infty}^{+\infty} f(t) \exp(-j2\pi\nu t) dt \tag{9.1}$$

while the inverse Fourier transform is defined by

$$\mathcal{F}^{-1}(\hat{f}(\nu)) = f(t) = \int_{-\infty}^{+\infty} \hat{f}(\nu) \exp(j2\pi\nu t) d\nu \tag{9.2}$$

Notice that passing from the Fourier transform to the Laplace transform is simply performed by changing $j2\pi\nu = j\omega$ into s in the integral term.

The main properties of the Fourier transformation are the following:
Linearity:
The Fourier transformation is a linear application:

$$\mathcal{F}(\alpha f(t) + \beta g(t)) = \alpha \hat{f}(\nu) + \beta \hat{g}(\nu) \tag{9.3}$$

Transformation sin-cos of the function $f(t)$:
Any function $f(t)$ can be decomposed into the sum of an even function $e(t)$ and an odd function $o(t)$

$$\begin{array}{rll} f(t) = & e(t) + & o(t) \\ & \text{even} & \text{odd} \end{array}$$

with

$$e(t) = \frac{1}{2}[f(t) + f(-t)] \qquad \text{and} \qquad o(t) = \frac{1}{2}[f(t) - f(-t)] \tag{9.4}$$

The Fourier transform of f is

$$\begin{array}{ll} \hat{f}(\nu) & = 2\displaystyle\int_0^{\infty} e(t)\ \cos(2\pi\nu t) dt - i2 \int_0^{\infty} o(t)\ \sin(2\pi\nu t) dt \\ & = \mathcal{F}_{\cos}(f(t)) - i\, \mathcal{F}_{\sin}(f(t)) \end{array} \tag{9.5}$$

Thus, the Fourier transform of function $f(t)$ is decomposed into a sum of two terms, one being the cosinus transform of the even part and the other being the sinus transform of the odd part.
Consequences:

$f(t)$	$\longrightarrow$ Fourier transform	$\hat{f}(\nu)$
even	$\longrightarrow$	even
odd	$\longrightarrow$	odd
real	$\longrightarrow$	hermitian (i.e.: $\hat{f}(\nu) = \overline{\hat{f}(-\nu)}$)
imaginary	$\longrightarrow$	skew-hermitian (i.e.: $\hat{f}(\nu) = -\overline{\hat{f}(-\nu)}$)

Transposition:

$$\mathcal{F}(f(-t)) = \hat{f}(-\nu) \tag{9.6}$$

Conjugate:

$$\mathcal{F}(\overline{f}(t)) = \overline{\hat{f}(-\nu)} \tag{9.7}$$

Scale change:

$$\mathcal{F}(f(at)) = \frac{1}{|a|}\hat{f}(\frac{\nu}{a}) \tag{9.8}$$

A compression of the timescale t induces a dilatation of the frequency scale (and reciprocally), due to the duality time-frequency.
Time translation:

$$\mathcal{F}(f(t-a)) \;=\; \exp(-j2\pi\nu a)\hat{f}(\nu) \tag{9.9}$$

The Fourier transform of the time-shifted function has its module unchanged, only its argument is modified.
Frequency modulation:

$$\mathcal{F}(\exp(j2\pi\nu_0 t)\, f(t)) = \hat{f}(\nu - \nu_0) \tag{9.10}$$

Modulating a function $f(t)$ by an imaginary exponential amounts to translating its Fourier transform.
Derivation with respect to variable t:
Assume that $f(t)$ is integrable, differentiable, with integrable derivative.

$$\mathcal{F}(f'(t)) = j2\pi\nu\, \mathcal{F}(f(t)) = j2\pi\nu\, \hat{f}(\nu) \tag{9.11}$$

and by iteration

$$\mathcal{F}(f^{(m)}(t)) = (j2\pi\nu)^m \mathcal{F}(f(t)) = (j2\pi\nu)^m \hat{f}(\nu) \tag{9.12}$$

The following inequality results

$$\int |f^{(m)}(t)|dt \geq |(2\pi\nu)^m|\, |\hat{f}(\nu)| \tag{9.13}$$

The more differentiable f is, with integrable derivatives, the faster $\hat{f}$ decreases towards infinity.
Derivation with respect to frequency ν:

$$\frac{d}{d\nu}\hat{f}(\nu) = \mathcal{F}((-j2\pi t)f(t)) \tag{9.14}$$

the previous derivation is justified if $tf(t)$ is integrable (thus $f(t)$ decreases faster than $1/t^2$).
More generally

$$\hat{f}^{(m)}(\nu) = \mathcal{F}((-j2\pi t)^m f(t)) \tag{9.15}$$

resulting in

$$|\hat{f}^{(m)}(\nu)| \leq \int |f(t)|\,|2\pi t|^m dt \tag{9.16}$$

the more $f(t)$ decreases towards infinity, the more differentiable $\hat{f}(\nu)$ is (with bounded derivatives).
Fourier transform of the convolution product:
the Fourier transform of the convolution product (denoted by *) of two functions is equal to the product of the Fourier transforms of both functions

$$\mathcal{F}(f(t) * g(t)) = \mathcal{F}(f(t))\ \mathcal{F}(g(t)) \tag{9.17}$$

Relation of Parseval-Plancherel:

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \int_{-\infty}^{+\infty} |\hat{f}(\nu)|^2 d\nu \tag{9.18}$$

It expresses that the total energy of signal $f(t)$ (represented by the left member) is equal to the sum of the energies of the signal components (right member).

Energy Properties

Total finite energy signals:
The functions $f(t)$ of the integrable square $|f(t)|^2$ play an important role in physics as, when t is the time variable, $|f(t)|^2$ represents, in general, an energy per time unit (power). The integral

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt \tag{9.19}$$

represents the total dissipated energy. In physics, the integrable-square functions are, in general, called total finite energy functions.

The mean quadratic deviation is defined by

$$< t^2 >= \frac{\int_{-\infty}^{+\infty} t^2 |f(t)|^2 dt}{\int_{-\infty}^{+\infty} |f(t)|^2 dt} \tag{9.20}$$

and represents an image of the concentration of the energy in the neighbourhood of 0.

The uncertainty relation (Roddier, 1971)

$$\sqrt{< t^2 >}\quad \sqrt{< \nu^2 >}\quad \geq\quad \frac{1}{4\pi} \tag{9.21}$$

means that if $\hat{f}(\nu)$ is searched for a given frequency ν_0, then one must seek in the past history and the future of the signal $f(t)$ for what corresponds to ν_0; this corresponds to an infinitely selective filtering which is impossible: $\hat{f}(\nu)$ cannot be perfectly known. Reciprocally, to know $f(t)$ from $\mathcal{F}(\nu)$, an infinite bandwidth is necessary.

By setting the Fourier transform of $f(t)$ equal to

$$\hat{f}(\nu) = A(\nu) \exp(j\alpha(\nu)) \tag{9.22}$$

the module $A(\nu)$ is positive real and the argument $\alpha(\nu)$ is also real. Thus,

$$|\hat{f}(\nu)|^2 = A^2(\nu) = \Phi_f^0(\nu) \tag{9.23}$$

The energy spectral density $\Phi_f^0(\nu)$ (also called energy spectrum) does not depend on the signal phase spectrum, thus is not modified by a translation of the signal on the time axis.

The autocorrelation function of the finite energy signal $f(t)$ is defined by

$$\phi_f^0(\xi) = \int_{-\infty}^{+\infty} f(t)\overline{f(t-\xi)}dt \tag{9.24}$$

where $\overline{f(x)}$ represents the conjugate complex function of $f(x)$. The Fourier transform of the autocorrelation function of the finite energy signal $f(t)$ is the energy spectral density $\Phi_f^0(\nu)$. The autocorrelation function of a periodic function of period T is also a periodic function of period T. The autocorrelation function can be normalized

$$\Gamma(\xi) = \frac{\phi_f^0(\xi)}{\phi_f^0(0)} \tag{9.25}$$

$\Gamma(\xi)$ is called the coherence or degree of self-coherence: $0 \leq \Gamma(\xi) \leq 1$.

The cross-correlation function of the finite energy signals $f(t)$ and $g(t)$ is defined by

$$\phi_{fg}^0(\xi) = \int_{-\infty}^{+\infty} f(t)\,\overline{g(t-\xi)}dt \tag{9.26}$$

The Fourier transform of the cross-correlation function of the finite energy signals $f(t)$ and $g(t)$ is the cross-spectral density $\Phi_{fg}^0(\nu)$. The cross-correlation function can also be normalized

$$\Gamma_{fg}(\xi) = \frac{\phi_{fg}^0(\xi)}{\sqrt{\phi_f^0(0)\,\phi_g^0(0)}} \tag{9.27}$$

the function $\Gamma_{fg}(\xi)$ is the coherence degree between f and g.

Finite mean power signals:

Some signals such as $f(t) = a\cos(\omega t)$ do not have a finite energy. In this case, the mean power transported by the signal is considered

$$\lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |f(t)|^2 dt \tag{9.28}$$

The previous properties are extended to finite mean power signals, e.g. the autocorrelation

$$\phi_f(\xi) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\,\overline{f(t-\xi)}dt \tag{9.29}$$

Periodical signals:
Let $f(t)$ be a periodic function of period T. In a general manner, it can be shown that $f(t)$ can be expanded as a series of orthogonal[1] functions in the considered interval $[t_0, t_0 + T]$

$$f(t) = \sum_{n=-\infty}^{n=+\infty} c_n \exp(j2\pi n \frac{t}{T}) \tag{9.31}$$

where the coefficients c_n are equal to

$$c_n = \frac{1}{T} \int_{t_0}^{t_0+T} f(t) \exp(-j2\pi n \frac{t}{T}) dt \tag{9.32}$$

When the function $f(t)$ is real, the Fourier series expansion is written as

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{n=+\infty} \left(a_n \cos(\frac{2\pi}{T} nt) + b_n \sin(\frac{2\pi}{T} nt) \right) \tag{9.33}$$

or

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{n=+\infty} (a_n \cos(2\pi\nu_0 nt) + b_n \sin(2\pi\nu_0 nt)) \tag{9.34}$$

by introducing $\nu_0 = 1/T$.

The coefficients a_n and b_n can be calculated according to the following relations

$$\begin{aligned} a_n &= \tfrac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \cos(2\pi\nu_0 nt) dt \\ b_n &= \tfrac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \sin(2\pi\nu_0 nt) dt \end{aligned} \tag{9.35}$$

Set: $\hat{f}_n = \hat{f}(n\nu_0) = \frac{1}{2}(a_n - jb_n)$. It can be noticed that the coefficient c_n previously defined is equal to $\hat{f}(n\nu_0)$. The Fourier transform $\hat{f}(\nu)$ of function $f(t)$ is then equal to

$$\hat{f}(\nu) = \sum_{n=-\infty}^{n=+\infty} \hat{f}(n\nu_0)\delta(\nu - n\nu_0) \tag{9.36}$$

[1]The functions $f(t)$ and $g(t)$ are defined as orthogonal in the interval $[a, b]$ if their scalar product is zero

$$< f, \bar{g} >= \int_a^b w(t)\, f(t)\overline{g(t)} dt = 0 \tag{9.30}$$

where $w(t)$ is a weight function $[a, b]$ over a particular interval, both depending on the considered type of function. Examples of orthogonal functions are Walsh functions (periodic, binary (= −1 and +1), called square waves in electronics), Legendre polynomials, Laguerre polynomials, Chebyshev polynomials and Hermite polynomials.

$\hat{f}(n\nu_0)$ is the frequency spectrum, which is decomposed into the amplitude spectrum

$$|\hat{f}(n\nu_0)| = \frac{1}{2}\sqrt{a_n^2 + b_n^2} \tag{9.37}$$

which is an even function, and its phase spectrum

$$\phi(n\nu_0) = \text{arctg}\left(-\frac{b_n}{a_n}\right) \tag{9.38}$$

which is an odd function.

The spectrum of a periodic function of period T is a discrete spectrum, thus a discrete function whose minimum interval on the frequency axis is: $\nu_0 = 1/T$; it is discontinuous (it exists only for multiples of the fundamental frequency ν_0).
Non periodic signals:
The non-periodicity of the studied function can be considered as the fact that when $T \to \infty$, then $\nu_0 \to 0$, the spectrum becomes a continuous function. The function $f(t)$ is deduced in a form analogous to that given for periodic functions

$$f(t) = \int_{-\infty}^{+\infty} \exp(j2\pi\nu t)d\nu \int_{-\infty}^{+\infty} f(u)\ \exp(-j2\pi\nu u)du \tag{9.39}$$

and the Fourier transform

$$\hat{f}(\nu) = \int_{-\infty}^{+\infty} f(t)\ \exp(-j2\pi\nu t)dt \tag{9.40}$$

Fourier transform of signals known on a limited interval:
Here, only physical signals are concerned. As they are known in a limited interval, these signals have necessarily a finite energy.

Two cases are often met:
a) The function represented in known interval $[0, T]$ as the considered signal is assumed to be zero outside, thus its Fourier transform $\hat{f}(\nu)$ is a continuous function with respect to ν.
b) The function represented on known interval $[0, T]$ as the considered signal is assumed periodic with period T. Its Fourier transform $\hat{f}(\nu)$ is a discrete spectrum with rays $\Delta\nu = 1/T$, each ray being equal to

$$\hat{f}_n = \hat{f}(\nu) \qquad \text{with: } \nu = n/T \tag{9.41}$$

The larger the period T, the larger the density of the discrete spectrum. The Fourier transform is equal to

$$\hat{f}(\nu) = \sum_{n=-\infty}^{n=+\infty} \hat{f}_n\, \delta(\nu - \frac{n}{T}) = \sum_{n=-\infty}^{n=+\infty} \hat{f}(\frac{n}{T})\, \delta(\nu - \frac{n}{T}) \tag{9.42}$$

9.1.2 Discrete Fourier Transform

Obtaining the discrete Fourier transform (Fig. 9.1) can be realized (Brigham, 1974) according to the following procedure

1. First, consider the continuous signal $f(t)$.
2. Sample this signal with a Dirac comb δ_{T_s} of period T_s (this is a periodic train of Dirac or ideal impulses), of amplitude 1; the sampled function is thus obtained
$$f(t)\,\delta_{T_s} = \sum_{n=-\infty}^{+\infty} f(nT_s)\delta(t-nT_s) \tag{9.43}$$
Note that a bad choice of the sampling period can generate spectral aliasing.
3. This function is truncated by multiplying by a rectangular pulse function $g(t)$ of length D, amplitude 1, not centred in 0, not starting at 0, but at $-T_s/2$ (this allows us to keep the signal at 0). N values are obtained such that
$$\begin{aligned} f(t)\,\delta_{T_s}\,g(t) &= \textstyle\sum_{n=-\infty}^{+\infty} f(nT_s)\,\delta(t-nT_s)\,g(t) \\ &= \textstyle\sum_{n=0}^{N-1} f(nT_s)\,\delta(t-nT_s) \end{aligned} \tag{9.44}$$
with $N = D/T_s$. We could consider that this sampled set of N values represents a pattern of a periodic function defined in $[-\infty, +\infty]$.
4. The Fourier transform of this sampled set of N values is equal to
$$\begin{aligned} \hat{f}(\nu) &= \mathcal{F}[\textstyle\sum_{n=0}^{N-1} f(nT_s)\,\delta(t-nT_s)] \\ &= \textstyle\sum_{n=0}^{N-1} f(nT_s)\,\mathcal{F}[\delta(t-nT_s)] \\ &= \textstyle\sum_{n=0}^{N-1} f(nT_s)\,\exp(-j2\pi\nu nT_s) \end{aligned} \tag{9.45}$$

The previous Fourier transform $\hat{f}(\nu)$ was a continuous function with respect to frequency ν, periodic (with its period equal to $\nu_s = 1/T_s$), completely defined by the N values of the series (9.45). The time-truncation by the rectangular pulse function causes frequency oscillations, called the Gibbs phenomenon (Kwakernaak and Sivan, 1991). This continuous Fourier transform is itself sampled at frequencies $\nu_k = k\Delta\nu$ (with $\Delta\nu = 1/(NT_s)$), called the harmonic frequencies of the discrete Fourier transform. The discrete Fourier transform (DFT) is defined (Kunt, 1981) as the series

$$\hat{f}(k) = \sum_{n=0}^{N-1} f(n)\,\exp(-j2\pi kn/N) \tag{9.46}$$

where $f(n)$ represents the value of function $f(t)$ at time nT_s. $\hat{f}(k)$ thus is an element of a periodic series of N elements defined in a frequency bandwidth $\nu_s = 1/T_s$ and separated by the frequency increment $\Delta\nu = 1/(NT_s)$.

The inverse discrete Fourier transform is defined as the series

$$f(n) = \frac{1}{NT_s} \sum_{k=0}^{N-1} \hat{f}(k)\,\exp(j2\pi nk/N) \tag{9.47}$$

Frequently, the sampling period is normalized: $T_s = 1$, resulting in slightly simplified formulae.

With the series $f(n)$ being periodic, it is possible to make it start at any instant, which gives

$$\hat{f}(k) = \sum_{n=n_0}^{n_0+N-1} f(n) \exp(-j2\pi kn/N) \tag{9.48}$$

In general, the time origin is chosen as the first instant, giving formula (9.46).

Similarly, as the series $\hat{f}(k)$ is periodic, it is possible to make it start at any frequency. Often, the main period of the complex spectrum is chosen as the interval $[-N/2, N/2-1]$, hence the inverse discrete Fourier transform

$$f(n) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \hat{f}(k) \exp(j2\pi nk/N) \tag{9.49}$$

The number of points N must be large and, in general, are taken as a power of 2 (e.g. $N = 1024$). In these conditions, the discrete Fourier transform is equivalent to the Fourier transform of the continuous signal. The fast Fourier transform (FFT) corresponds to a fast computing algorithm of the discrete Fourier transform.

Some properties of the discrete Fourier transform are:

- The discrete Fourier transform is a linear application.
- The discrete Fourier transform is an unitary operation[2], i.e. it keeps the scalar product

$$< f_1, f_2 >=< \hat{f}_1, \hat{f}_2 > \tag{9.50}$$

  This property induces the Parseval-Plancherel relation

$$\|f\|^2 = \|\hat{f}\|^2 \tag{9.51}$$

- The signal energy (cf. Parseval-Plancherel relation) is equal to

$$\sum_{n=0}^{N-1} f(n)^2 = \sum_{k=-N/2}^{N/2-1} |\hat{f}(k)|^2 \tag{9.52}$$

  i.e. the sum of the energies of the signal frequency components $\nu_k = k/(NT_s)$. Each frequency contribution $|\hat{f}(k)|^2$ is sometimes called a periodogram (Ljung, 1987), and constitutes an estimation of the spectral density, however biased (Söderström and Stoica, 1989).
- The discrete convolution product y of two signals u and g is defined by

$$y(n) = \sum_{i=0}^{N-1} u(i)g(n-i) \ , \quad n = 0, \ldots, N-1 \tag{9.53}$$

[2] A linear transformation defined in a vector space is unitary if it keeps the scalar product and the norm.

- The discrete Fourier transform of the convolution product is equal to the product of the discrete Fourier transforms

$$\hat{y}(k) = \hat{u}(k)\hat{g}(k) \tag{9.54}$$

- The time translation m induces a phase rotation of the discrete transform

$$\hat{f}_m(k) = \sum_{n=0}^{N-1} f(n-m)\, \exp(-j2\pi kn/N) = \hat{f}(k) \exp(-j2\pi km/N) \tag{9.55}$$

- The discrete Fourier transform verifies

$$\hat{f}(-k) = \overline{\hat{f}(k)} \tag{9.56}$$

Remark:
Kwakernaak and Sivan (1991) call the discrete Fourier transform previously defined "Discrete to discrete Fourier transform", to emphasize that the discrete signal $f(n)$ is used to build the discrete signal $\hat{f}(k)$. This transform uses an expansion on an orthogonal basis. By using an expansion on an orthonormal basis with respect to the scalar product, they define, as Ljung (1987), the discrete Fourier transform as

$$\hat{f}(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} f(n)\, \exp(-j2\pi kn/N) \quad , \quad k = 0, \ldots, N-1 \tag{9.57}$$

and the inverse discrete Fourier transform

$$f(n) = \frac{1}{\sqrt{N}} \sum_{k=-N/2}^{N/2-1} \hat{f}(k)\, \exp(j2\pi nk/N) \quad , \quad n = 0, \ldots, N-1 \tag{9.58}$$

Moreover, Kwakernaak and Sivan (1991) define the sampled discrete Fourier transform as

$$\hat{f}(\nu) = T_s \sum_{t=0}^{(N-1)T_s} f(t)\, \exp(-j2\pi\nu t) \quad , \quad \nu = 0, \ldots, (N-1)/(NT_s) \tag{9.59}$$

and the inverse sampled discrete Fourier transform as

$$f(t) = \frac{1}{NT_s} \sum_{\nu=0}^{(N-1)/(NT_s)} \hat{f}(\nu)\, \exp(j2\pi\nu t) \quad , \quad t = 0, \ldots, (N-1)T_s \tag{9.60}$$

These various transforms are, of course, also unitary.

9.1.3 Stochastic Signals

A signal is nearly always noisy. The noise is a stochastic (random) phenomenon which is superposed to the source signal of interest to the user. The noise and

the noisy signal constitute two stochastic signals. In continuous time, the signal is denoted by $x(t)$. In discrete time, the signal is denoted by $x(n)$, where n is the instant. With the stochastic signal $x(n)$ is associated a distribution function $f(x, n)$ equal to

$$f(x, n) = \mathcal{P}(x(n) \leq x) \tag{9.61}$$

where $\mathcal{P}(x(n) \leq x) = \mathcal{P}(x, n)$ is the probability that the variable $x(n)$ is lower than a given value x. The probability density $p(x, n)$ is the derivative of the distribution function, defined as

$$p(x, n) = \frac{df(x, n)}{dx} \tag{9.62}$$

The main statistical variables characterizing the random signal $x(n)$ are:

- The first-order moment or set mean or mathematical expectation of the signal resulting from a large number of observations

$$\mu_x(n) = E[x(n)] = \int_{-\infty}^{\infty} x(n)p(x, n)dx(n) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i(n)\, \mathcal{P}(x, n) \tag{9.63}$$

- The time mean of the signal

$$\mu_x = \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t)dt = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x(i) \tag{9.64}$$

- The second-order moment

$$E[x^2(n)] = \int_{-\infty}^{\infty} x^2(n)p(x, n)dx(n) \tag{9.65}$$

- The second-order moment centred with respect to the mean, or variance

$$\sigma_x^2(n) = E[\{x(n) - \mu_x(n)\}^2] = \int_{-\infty}^{\infty} [x(n) - \mu_x(n)]^2 p(x, n)dx \tag{9.66}$$

When the stochastic signal x is considered at two different instants n_1 and n_2, the distribution function of the stochastic variables $x(n_1)$ and $x(n_2)$ is defined by

$$f(x_1, x_2, n_1, n_2) = \mathcal{P}(x(n_1) \leq x_1; x(n_2) \leq x_2) \tag{9.67}$$

where $\mathcal{P}(x(n_1) \leq x_1; x(n_2) \leq x_2) = \mathcal{P}(x_1, x_2)$ is the joint probability (or second-order probability) that the signal x at time n_1 is lower than a given value x_1, and that x at n_2 is lower than a given value x_2. The joint probability density results

$$p(x_1, x_2, n_1, n_2) = \frac{\partial^2 f(x_1, x_2, n_1, n_2)}{\partial x_1 \partial x_2} \tag{9.68}$$

The associated statistical properties are:

- The autocorrelation function (denoted by ϕ or R) of the signal $x(n)$

$$R_{xx}(n_1, n_2) = E[x(n_1)x(n_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 p(x_1, x_2, n_1, n_2)dx_1 dx_2 \tag{9.69}$$

- The autocovariance function of the signal $x(n)$

$$\begin{aligned} C_{xx}(n_1, n_2) &= E[\{x(n_1) - \mu_x(n_1)\}\{x(n_2) - \mu_x(n_2)\}] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x_1 - \mu_x(n_1)][x_2 - \mu_x(n_2)]p(x_1, x_2, n_1, n_2)dx_1 dx_2 \\ &= R_{xx}(n_1, n_2) - \mu_x(n_1)\mu_x(n_2) \end{aligned} \tag{9.70}$$

9.1.4 Stochastic Stationary Signals

A signal is stationary in the strict sense if all its statistical properties are time-independent. It is stationary in the wide sense if its first two moments are time-independent. The notations are then simplified:

- The probability density, the mean, the variance are, respectively, $p(x)$, μ_x, σ_x^2.
- Noting that $\tau = t_2 - t_1$, the autocorrelation function and the autocovariance function are, respectively $R_{xx}(\tau)$ and $C_{xx}(\tau) = R_{xx}(\tau) - \mu_x^2$.

The autocorrelation function does not depend anymore on time t, but only on the time shift τ.

A signal is ergodic when the mean values (time averages) are equivalent to the corresponding expectations (set averages), thus for a moment of any order i

$$E[x^i] = \int_{-\infty}^{\infty} x^i p(x)dx \equiv \overline{x^i} = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} x^i(n) \tag{9.71}$$

Thus, the autocorrelation function of a stationary and ergodic signal is equal to

$$R_{xx}(n) = E[x(i)x(i+n)] = \overline{x(i)x(i+n)} = \frac{1}{N-|n|}\sum_{i=0}^{N-|n|-1} x(i)x(i+n) \tag{9.72}$$

and the autocovariance function of a stationary and ergodic signal is equal to

$$C_{xx}(n) = \overline{[x(i) - \mu_x][x(i+n) - \mu_x]} = R_{xx}(n) - \mu_x^2 \tag{9.73}$$

Recall that the power spectral density Φ_x of a signal x is the Fourier transform of the autocorrelation function of this signal

$$\Phi_x(k) = \sum_{n=-\infty}^{\infty} R_{xx}(n)\exp(-j2\pi kn/N) \tag{9.74}$$

The cross-correlation function of the stationary and ergodic signals x and y is equal to

$$R_{xy}(n) = E[x(i)y(i+n)] = \overline{x(i)y(i+n)} = \frac{1}{N-|n|} \sum_{i=0}^{N-|n|-1} x(i)y(i+n) \quad (9.75)$$

The two signals are not correlated if the cross-correlation is always zero. The cross-spectral density Φ_{xy} of the signals x and y is the Fourier transform of the cross-correlation function of these signals

$$\Phi_{xy}(k) = \sum_{n=-\infty}^{\infty} R_{xy}(n) \exp(-j2\pi kn/N) \quad (9.76)$$

The cross-covariance function of the stationary and ergodic signals x and y is equal to

$$C_{xy}(n) = E[\{x(i) - \mu_x\}\{y(i+n) - \mu_y\}] = R_{xy}(n) - \mu_x\mu_y \quad (9.77)$$

9.1.5 Summary

Consider a variable such as signal energy. The expression of this variable differs according to the type of signal that is considered and whether time is continuous or discrete. Table 9.1 resumes the different cases that can be encountered. It is possible to have the same reasoning for other quadratic variables such as power, scalar product, correlation (see Table 9.2), covariance, spectral density, etc.

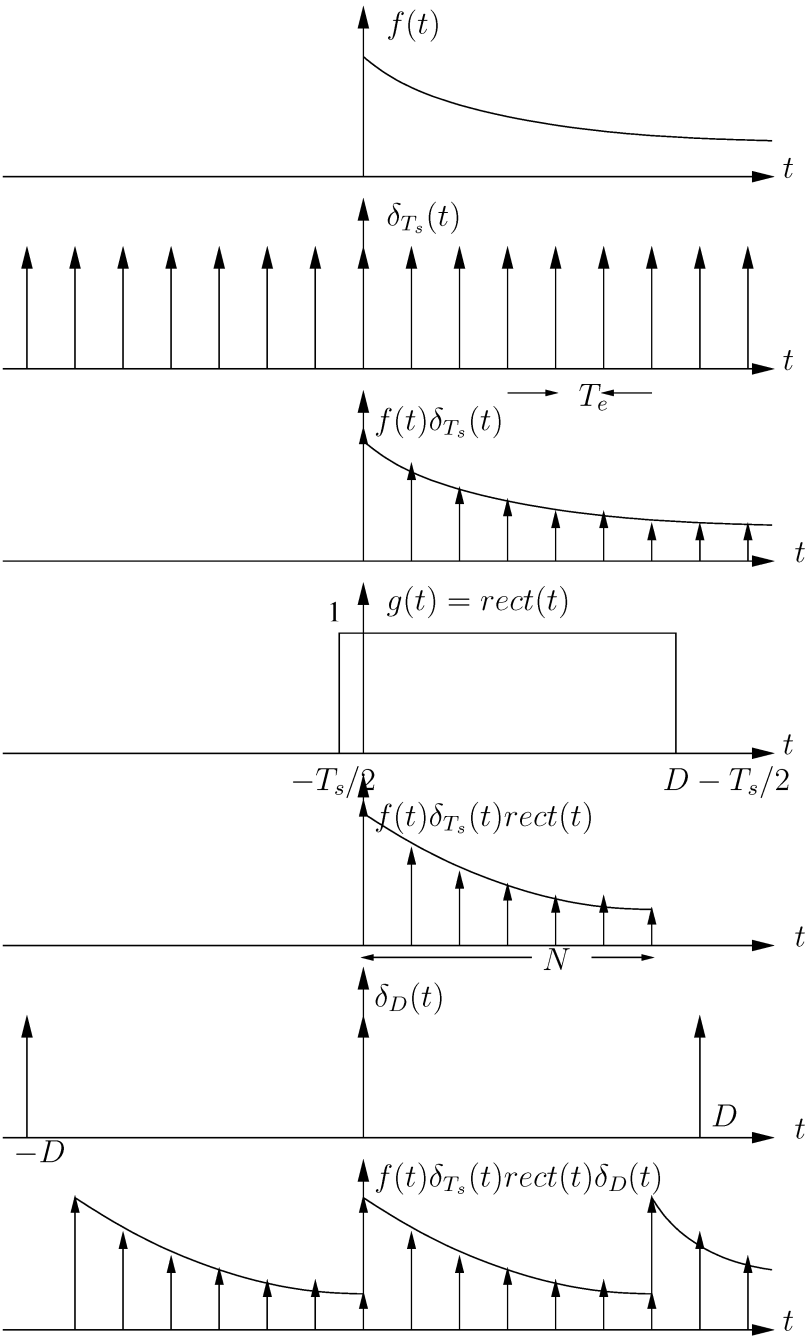


Figure 9.1: Obtaining the discrete Fourier transform

Table 9.1: Obtaining a variable such as energy or power according to the type of signal and continuous or discrete time

Type of signal $x(t)$	Continuous time	Discrete time
Finite energy	$E = \int_{-\infty}^{+\infty} \lvert x(t)\rvert^2 dt$	$E = \sum_{i=-\infty}^{+\infty} \lvert x(i)\rvert^2$
Finite mean power, periodic signal	$P = \frac{1}{T}\int_{t_0}^{t_0+T} \lvert x(t)\rvert^2 dt$ period T	$P = \frac{1}{N}\sum_{i=n_0}^{n_0+N} \lvert x(i)\rvert^2$ period N
Finite mean power, non-periodic signal	$P = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \lvert x(t)\rvert^2 dt$	$P = \lim_{N\to\infty} \frac{1}{2N+1}\sum_{i=-N}^{N} \lvert x(i)\rvert^2$
Finite mean power, random signal	$P = E[\lvert(x(t) - \mu_x)\rvert^2]$	$P = E[\lvert(x(n) - \mu_x)\rvert^2]$

Table 9.2: Obtaining the correlations (auto if $x = y$ or cross if $x \neq y$) according to the type of signal and continuous or discrete time

Signal $x(t)$	Continuous time	Discrete time
Finite energy	$\phi_{xy}(\tau) = \int_{-\infty}^{+\infty} x(t)y^*(t-\tau)dt$	$\phi_{xy}(m) = \sum_{i=-\infty}^{+\infty} x(i)y^*(i-m)$
Finite mean power, periodic signal	$\phi_{xy}(\tau) = \frac{1}{T}\int_{t_0}^{t_0+T} x(t)y^*(t-\tau)dt$ period T	$\phi_{xy}(m) = \frac{1}{N}\sum_{i=n_0}^{n_0+N} x(i)y^*(i-m)$ period N
Finite mean power, non-periodic signal	$\phi_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} x(t)y^*(t-\tau)dt$	$\phi_{xy}(m) = \lim_{N\to\infty} \frac{1}{2N+1}\sum_{i=-N}^{N} x(i)y^*(i-m)$
Finite mean power, random signal	$\phi_{xy}(\tau) = E[(x(t)-\mu_x)(y(t-\tau)-\mu_y)]$	$\phi_{xy}(m) = E[(x(n)-\mu_x)(y(n-m)-\mu_y)]$

9.2 Sampling

9.2.1 D/A and A/D Conversions

Suppose that the signal delivered by the sensor is analog. This signal is not usable by the computer and must be transformed: using a clock, a given number of signal values are withdrawn, thus the analog signal is sampled with a sampling period Δt (Fig. 9.2) or a sampling frequency $1/\Delta t$. The signal thus obtained consists of a sequence of discrete values separated by a time interval Δt (impulse modulation). The sampling operation induces a loss of information: if the sampling is regular, no information exists between two successive equidistant instants. A multiplexer allows the acquisition of several signals coming from different sensors, of which only a small proportion will really be used for control, the remaining being used for monitoring. Note that, in practice, the sampling is frequently realized either with different sampling periods according to the signals, or with time-varying sampling periods.

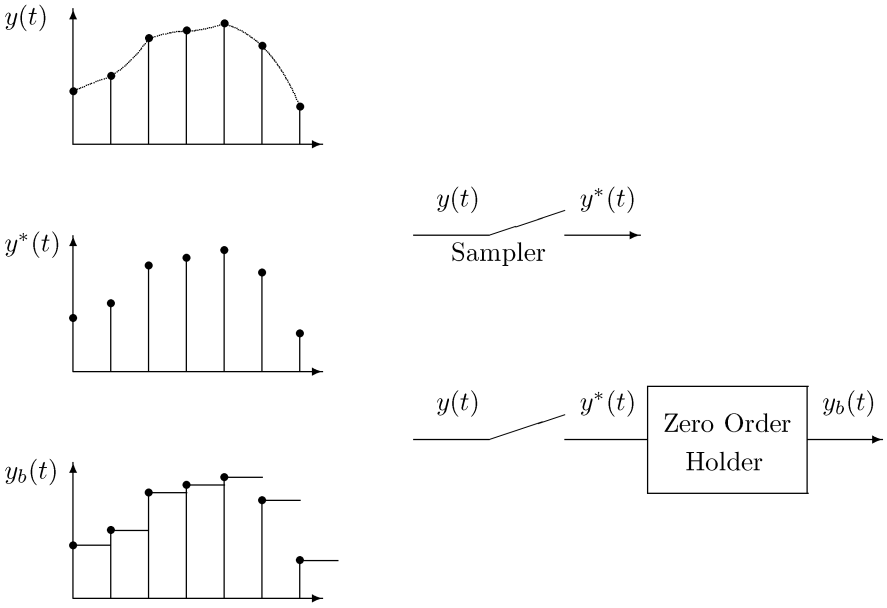


Figure 9.2: Sampling of a signal

Conversely, if the original signal is a digital signal delivered by the computer and is to be transformed into an analog signal, a holder is used, e.g. a zero-order holder, also called a boxcar which keeps the signal value constant during the period Δt

$$y_b(t) = y_{n-1} = y(t_{n-1}) \quad \text{for} \quad t_{n-1} \leq t < t_n \tag{9.78}$$

The digital signal (Fig. 9.2), consisting of a series of impulses is thus discrete; it becomes continuous after the holder (signal reconstruction), and consists of

a series of steps.

Other types of holder can be used, such as the first-order holder, which linearly extrapolates the signal in the period $[t_{n-1}, t_n]$ by using the increase during the previous period

$$y_b(t) = y_{n-1} + \frac{(t - t_{n-1})}{\Delta t}(y_{n-1} - y_{n-2}) \quad \text{for } t_{n-1} \leq t < t_n \tag{9.79}$$

In fact, the more frequently used holder in process control remains the zero-order holder and we will assume that we always use this type. The general representation of the digital control loop is given by Fig. 9.3.

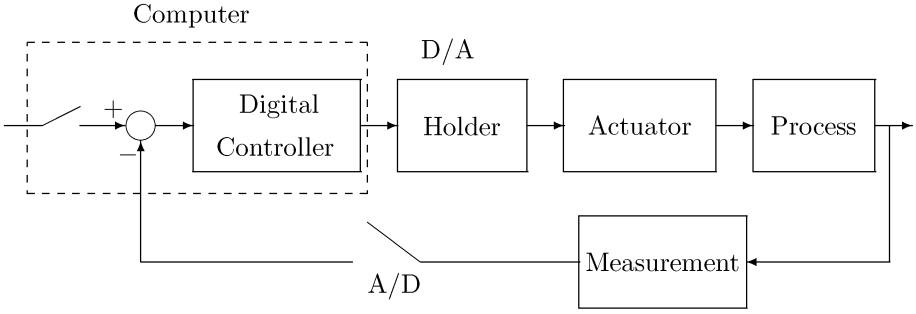


Figure 9.3: General representation of the digital control

9.2.2 Choice of Sampling Period

Consider the case of a sinusoidal signal (Fig. 9.4): we notice that if the sampling period is equal, for example, to the period of the sinusoid itself, the sampled signal contains information which will all have the same value, as if it were coming from a flat constant signal. This phenomenon is due to the sampling period being too large with respect to the frequency band of interest for the observed signal.

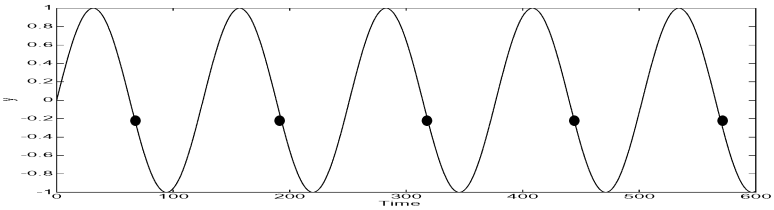


Figure 9.4: Sampled sinusoidal signal with an ill-chosen sampling period

A sampler is obtained (Fig. 9.5) by multiplying the input signal $y(t)$ by a sampling function $s(t)$, which is a periodic sequence (period $T_s = 1/\nu_s$) of rectangular impulses of duration D. The sampled signal is thus equal to

$$y_s(t) = y(t)\, s(t) \tag{9.80}$$

Imagine that the impulse duration D is infinitely short; the sampling function $s(t)$ then consists of a periodic train of ideal impulses (period T_s) denoted by δ_{T_s}

$$s(t) \;=\; \delta_{T_s} \;=\; \sum_{n=-\infty}^{n=+\infty} \delta(t - nT_s) \tag{9.81}$$

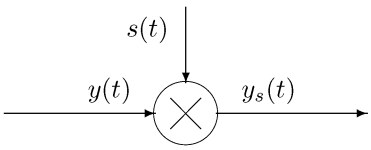


Figure 9.5: Sampling of a signal by a periodic train of ideal impulses

The ideally sampled signal $y_s(t)$ can be represented by a series

$$y_s(t) \;=\; \sum_{n=-\infty}^{n=+\infty} y(nT_s)\delta(t - nT_s) \tag{9.82}$$

The Fourier transform of the sampled function $y_s(t)$ is equal to the convolution product of the Fourier transform of the signal $y(t)$ by the Fourier transform of the sampling function $s(t)$

$$\hat{y_s}(\nu) = \hat{y}(\nu) * \hat{s}(\nu) \tag{9.83}$$

The Fourier transform of the periodic train of ideal impulses is equal to

$$\hat{\delta_{Ts}}(\nu) = \nu_s \delta_{\nu_s}(\nu) \tag{9.84}$$

which is itself a periodic train of ideal impulses, of period ν_s. The Fourier transform of the sampled function $y_s(t)$ results

$$\begin{aligned} \hat{y_s}(\nu) \quad &= \quad \hat{y}(\nu) * \nu_s \delta_{\nu_s}(\nu) \\ &= \quad \textstyle\sum_{n=-\infty}^{n=+\infty} \nu_s \hat{y}(\nu - n\nu_s) \end{aligned} \tag{9.85}$$

This function $\sum_{n=-\infty}^{n=+\infty} \nu_s \hat{y}(\nu - n\nu_s)$ is a periodic function of period ν_s and amplitude $\nu_s|\hat{y}(\nu)|$. The spectral density $\Phi^0_{ys}(\nu)$ (Fig. 9.6) is equal to

$$\Phi^0_{ys}(\nu) = \sum_{n=-\infty}^{n=+\infty} \nu_s^2 \Phi^0_y(\nu - n\nu_s) \tag{9.86}$$

Thus, it has the same period as the Fourier transform $\hat{y_s}(\nu)$. The functions $\hat{y_s}(\nu)$ and $\Phi^0_{ys}(\nu)$ are, respectively, the repetition of the motives of Fourier transform $\hat{y}(\nu)$ and spectral density $\Phi^0_y(\nu)$, respectively multiplied by ν_s and ν_s^2. The frequency $\nu_{\max}$ of Fig. 9.6 corresponds to the bandwidth of the process i.e. the frequency above which the amplitude of the spectral density becomes negligible.

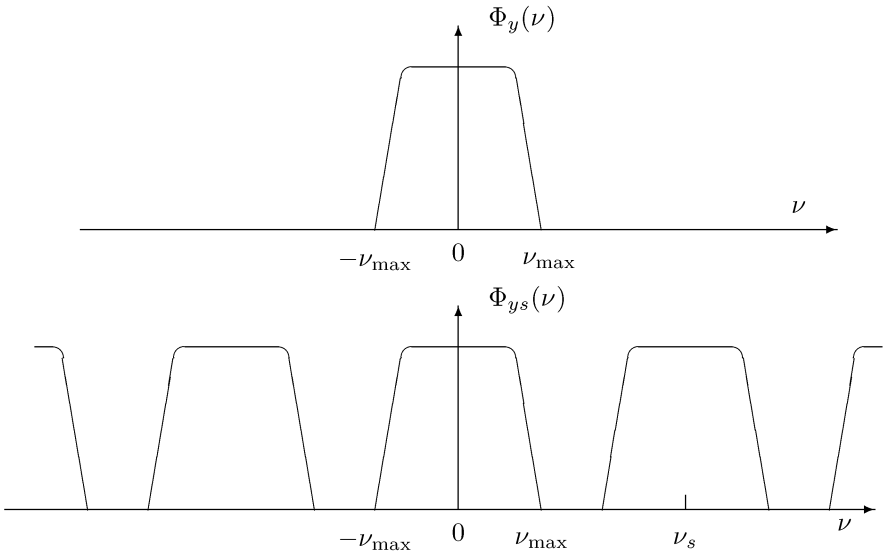


Figure 9.6: Spectral densities for the studied signal and the corresponding ideally sampled signal

This case was an ideal view; in reality, the impulse has a finite duration and the value obtained during duration D is held, or rather, an average is realized in this time interval. Thus, the first value obtained at $t = 0$ will be kept during D and the first rectangular impulse is centred at $D/2$; the following impulse starts after a period T_s and has the same duration D, and so on (Fig. 9.7).

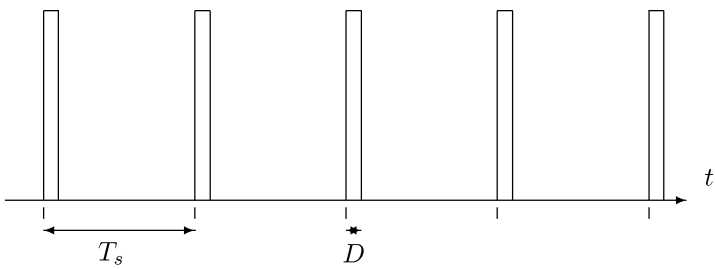


Figure 9.7: Real sampling of a signal

In a similar manner to the ideal case, the Fourier transform of the sampled function $y_s(t)$ is to be calculated. The Fourier transform of a rectangular impulse function of duration D, centred at t_0 (here: $D/2$) is

$$\mathcal{F}[\text{rect}(\frac{t-t_0}{D})] = D\text{sinc}(\nu D)\exp(-j2\pi\nu t_0) \tag{9.87}$$

where sinc represents the cardinal sinus function equal to $\text{sinc}(x) = \sin(\pi x)/(\pi x)$.

The Fourier transform of the sampled signal is

$$\begin{aligned} \hat{y_s}(\nu) &= \hat{y}(\nu) * [\textstyle\sum_{n=-\infty}^{n=+\infty} \nu_s D \mathrm{sinc}(nD\nu_s) \exp(-j\pi\nu D)] \\ &= \textstyle\sum_{n=-\infty}^{n=+\infty} \nu_s D \mathrm{sinc}(nD\nu_s) \hat{y}(\nu - n\nu_s) \exp(-j\pi\nu D) \end{aligned} \tag{9.88}$$

We thus notice that the Fourier transform of the sampled signal depends on the sampling frequency ν_s; it is no more a periodic function, but is attenuated by the presence of the sinus cardinal term. The exponential term is the consequence of the delay effect introduced by $D/2$.

The spectral density of the sampled signal $\Phi_{ys}^0(\nu)$ is not limited with respect to frequency, but extends in the whole spectrum (Fig. 9.8).

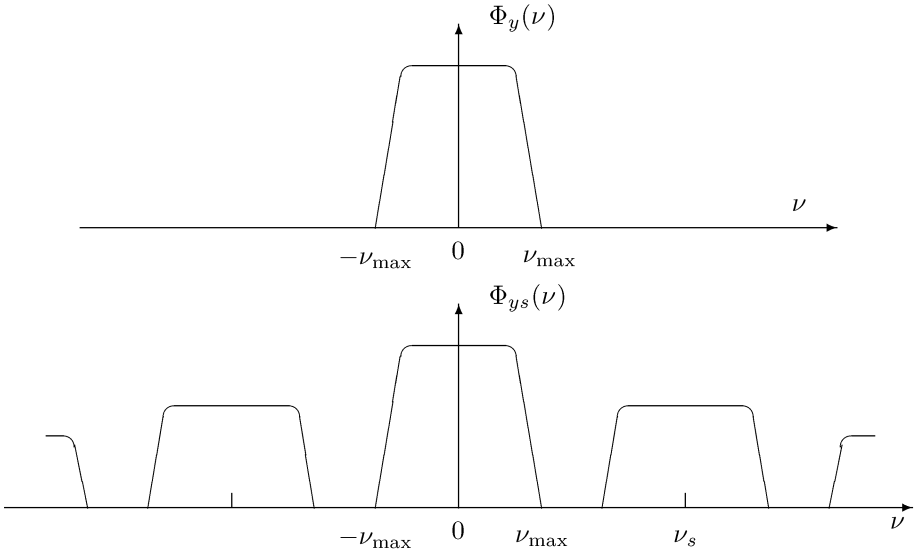


Figure 9.8: Spectral densities for the studied signal and the corresponding sampled signal

The original spectrum $\hat{y}(\nu)$ of the analog signal is limited at the maximum frequency $\nu_{\max}$. The spectrum of the sampled signal $\hat{y_s}(\nu)$ is a function (term: $\nu_s D \mathrm{sinc}(nD\nu_s) \exp(-i\pi\nu D)$) of the periodic repetition (term: $\sum_{n=-\infty}^{n=+\infty} \hat{f}(\nu - n\nu_s)$) of the original spectrum $\hat{y}(\nu)$ of the analog signal. If the term of the function were equal to 1, then the spectrum would be strictly periodic. According to the value of the maximum frequency ν_{max} corresponding to the extent of the spectrum of the analog signal, the sequences of the sampled signal spectrum either partly overlap themselves (Fig. 9.9) or do not overlap. If overlapping, also called aliasing, occurs, then the transformation is non-reversible, thus the reconstruction is not possible, which means a loss of information.

We assumed that the original spectrum of the analog signal was not frequency-limited; if that were not the case, frequency aliasing would occur for any sampling, even with ideal sampling. To avoid this problem, in practice, a

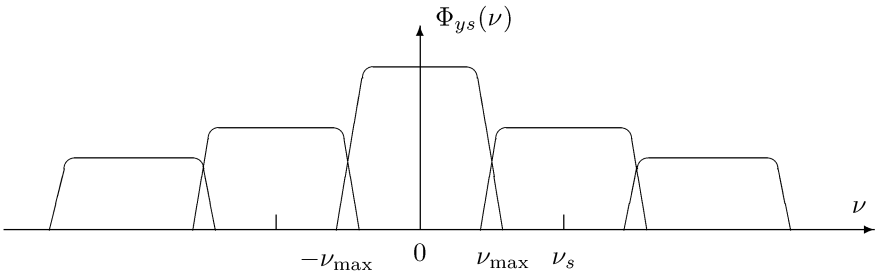


Figure 9.9: Frequency overlapping of the spectral densities: aliasing

convenient filter such as a high-order Butterworth filter is placed upwards with respect to the sampler in order to eliminate the higher frequencies.

A physically realizable signal (finite-energy signal) necessarily has an infinite extent spectrum. Thus, this would pose a problem concerning the sampling; but, so that the energy is finite, the spectrum must tend towards 0 when the frequency ν tends towards infinity. Thus, the spectrum will be nearly zero above some frequency; to reconstruct the signal as close as possible to $y(t)$, the high-frequency components of $\hat{y_s}(\nu)$ must be eliminated by a lowpass filter. The aliasing will be negligible provided that the sampling frequency is well chosen as the Shannon theorem indicates.

Shannon sampling theorem:

The Shannon theorem provides rules which allow us to guarantee minimal information losses caused by sampling: the analog signal $y(t)$, having a lowpass spectrum (band-limited signal) (maximum frequency $\nu_{\max}$), is totally described by the sequence of instantaneous values $y(t_k)$ periodically sampled provided that the sampling period T_s is lower than or equal to $1/(2\nu_{max})$

$$T_s \quad \leq \quad \frac{1}{2\nu_{max}} \qquad \text{or} \qquad \nu_s \quad \geq \quad 2\nu_{max} \tag{9.89}$$

In practice, the sampling frequency ν_s must be largely greater than ν_{max}, e.g.

$$\nu_s \quad \geq \quad 8\nu_{max} \tag{9.90}$$

In the case of an ideal sampler, the ideal lowpass filter would be a filter whose amplitude of the harmonic response is the rectangular function

$$|G(\nu)| \; = \; T_s \text{rect}(\frac{\nu}{\nu_s}) \tag{9.91}$$

When the spectrum of the signal to be sampled has a wide band due to the environment noise (sensors, electronics, etc.), a prefilter is placed before the sampler; this is called an anti-aliasing filter (Fig. 9.10), the function of which is to filter any frequency higher than $\nu_s/2$. Ideally, it would be a lowpass filter limited to a passband B (theoretically equal to the frequency ν_{max}), indeed practically extending up to B_m. The difference $B_m - B$ constitutes the transition band. Max (1985) recommends as an anti-aliasing filter the elliptic

Cauer filter, of order $n = 2m$, composed of m elementary filters in series, each having as a transfer function

$$G(s) = \frac{1 + \dfrac{s^2}{(2\pi\nu_0)^2}}{1 + 2\zeta \dfrac{s}{2\pi\nu_1} + \dfrac{s^2}{(2\pi\nu_1)^2}} \tag{9.92}$$

where ν_0 and ν_1 are characteristic frequencies. The Butterworth filter (Mitra and Kaiser, 1993) is frequently cited, but its transition from bandpass to bandstop is less stiff.

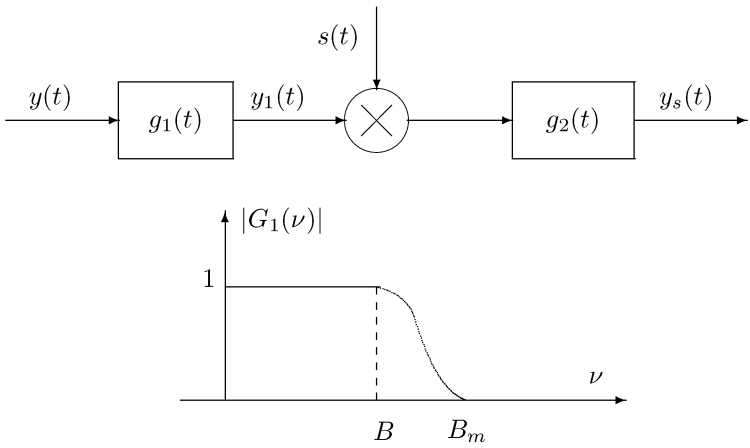


Figure 9.10: Anti-aliasing filtering

The sampling theorem cannot be applied without looking at the type of signal spectrum; that theorem was valid only for a lowpass signal; for a bandpass signal, the theorem can be expressed slightly differently. Suppose that the signal spectrum is bounded by low frequency ν_{min} and by high frequency ν_{max}; the bandwidth B is equal to $B = \nu_{max} - \nu_{min}$. Let m be the largest integer that is lower than or equal to ν_{max}/B.

To guarantee non-spectral overlapping, the sampling frequency must be larger than or equal to $2\nu_{max}/m$

$$\nu_s \quad \geq \quad \frac{2\nu_{max}}{m} \tag{9.93}$$

It is possible to notice that if the higher frequency ν_{max} is much larger than the bandwidth B, the minimum sampling frequency becomes close to $2B$

$$\text{If} \quad \nu_{max} \quad \gg \quad B \quad \Longrightarrow \quad \nu_s \quad \geq \quad 2B \tag{9.94}$$

Now we will discuss the practical choice of sampling period with respect to the real process characteristics:

- Too slow a sampling will have the drawback of reducing the efficiency of the feedback with respect to disturbances.
- Too fast a sampling will needlessly overload the computer memory. Moreover, when the signal to noise ratio (ratio of the signal variance over the variance of a random input disturbance or of a noise) is low, oversampling provokes a large influence of the noise: then it is necessary to filter.

Table 9.3: Recommendations for the choice of sampling period

Physical system or physical variable	Sampling period T_s recommended
Electrical engines	$T_s < 0.1$ s
Flow rate	1 s
Level	5 s
Pressure	5 s
Temperature	20 s
Open-loop system: characteristic variable	**Recommended sampling period**
First-order with time constant τ	$0.25\tau < T_s < \tau$
First-order with time delay t_d	$0.2t_d < T_s < t_d$
Second-order of natural frequency ω_n	$0.05/\omega_n < T_s < 1/\omega_n$
$\tau_{\max}$: dominant time constant	$T_s < 0.1\tau_{\max}$
Settling time t_s	$t_s/15 < T_s < t_s/6$
Critical frequency ω_c	$0.15/\omega_c < T_s < 0.5/\omega_c$
Closed-loop system: Characteristic variable	**Recommended sampling period**
Integral time constant τ_I	$T_s > \tau_i/100$
Derivative time constant τ_d	$0.1\tau_d < T_s < 0.5\tau_d$

Table 9.3 (Flaus, 1994; Seborg et al., 1989), gives some indications concerning the recommended sampling frequencies which are valid for a large number of chemical processes. However, these values may differ according to the process, especially concerning the cited physical variables, where the user will be essentially guided by the process time constants.

9.3 Filtering

The noise which affects analog signals can come from different sources: the measurement device, the electrical environment, the process, etc. The electrical noise can be minimized by shielding the cables, but the noises coming from the measurement and process must be minimized by filtering (Fig. 9.11), which

transforms a noisy input signal $y(t)$ into a filtered signal $y_f(t)$ coming out of the filter. A basic analog filtering is recommended, i.e. before data acquisition, where the signal can be later digitally filtered.

Figure 9.11: Filtering

Recall that the choice of sampling period is essential; to avoid the problems related to a too-low sampling rate, data signals are prefiltered by means of an anti-aliasing filter which will be used systematically in order to filter high-frequency noise.

It is not possible to present all existing filters, but rather to show some simple principles concerning elementary filters; for more complex filters (Butterworth, Chebyshev, Bessel, Cauer, ... , adaptive filters), the reader can refer to the following texts: Crochiere and Rabiner (1983); Kunt (1981); Mitra and Kaiser (1993); Proakis and Manolakis (1996); Rorabaugh (1997); Salman and Solotareff (1982). In particular, the transformation formulae from a given analog filter into a digital filter of specified bandwidth will be found (lowpass, bandpass, bandstop, highpass, notch). Bellanger (1989); Haykin (1991); Treichler et al. (1987); Vaidyanathan (1993) treat adaptive filters.

9.3.1 First-Order Filter

A first-order exponential filter will make a noisy measurement signal smoother. The differential equation corresponding to the first-order transfer function for the first-order analog filter is

$$\tau_f \frac{dy_f}{dt} + y_f(t) = y(t) \tag{9.95}$$

This filter is represented by a continuous first-order transfer function equal to

$$G_f(s) = \frac{Y_f(s)}{Y(s)} = \frac{1}{\tau_f s + 1} \tag{9.96}$$

where τ_f is the filter time constant and the filter gain is equal to 1. This filter dampens the high-frequency fluctuations: it is a lowpass filter currently called RC filter because of its electrical scheme. The filter time constant τ_f must be far smaller than the dominant time constant of the process $\tau_{\max}$ to avoid a lag in the feedback loop (e.g. $\tau_f < 0.1\tau_{\max}$). If the noise amplitude is large, τ_f must be increased. On the other hand, the noise bandwidth must be considered: let ν_m be the lowest frequency of the noise. τ_f must be chosen so that $\nu_f < \nu_m$ where $\nu_f = 1/\tau_f$ (the filter dampens signals of frequency higher than ν_f, i.e. the high-frequency noise). By defining $\nu_{\max} = 1/\tau_{\max}$, it results that for the filter

$$\nu_{\max} < \nu_f < \nu_m \tag{9.97}$$

Figure 9.12: First-order filter

It is possible to consider a digital first-order filter associated with Eq. (9.95), for example, by discretizing this equation using a backward finite difference scheme. Denote by subscript n the values at time t, by $y(n)$ the filter input and by $y_f(n)$ the filtered output (Fig. 9.12). The discretized differential equation becomes

$$\tau_f \frac{y_f(n) - y_f(n-1)}{\Delta t} + y_f(n) = y(n) \tag{9.98}$$

which can be ordered as

$$y_f(n) = \alpha y(n) + (1-\alpha) y_f(n-1) \qquad \text{with: } y_f(0) = y(0) \tag{9.99}$$

where

$$\alpha = \frac{1}{\tau_f/\Delta t + 1} \quad (0 < \alpha \leq 1) \tag{9.100}$$

The limit $\alpha = 1$ corresponds to no filtering (zero time constant). The limit $\alpha = 0$ does not take into account the measurement.

9.3.2 Second-Order Filter

The second-order filter is equivalent to two first-order filters in series. Let $y_{f2}(n)$ be the signal coming out from the second filter and $y_{f1}(n)$ from the first filter (Fig. 9.13).

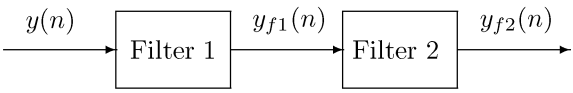


Figure 9.13: Two first-order filters in series

The equations for both filters are

$$y_{f1}(n) = \alpha_1 y(n) + (1-\alpha_1) y_{f1}(n-1) \tag{9.101}$$

$$y_{f2}(n) = \alpha_2 y_{f1}(n) + (1-\alpha_2) y_{f2}(n-1) \tag{9.102}$$

giving the equation for the set of filters

$$y_{f2}(n) = \alpha_1\alpha_2 y(n) + (2-\alpha_1-\alpha_2) y_{f2}(n-1) - (1-\alpha_1)(1-\alpha_2) y_{f2}(n-2) \tag{9.103}$$

which corresponds to a second-order filter. The second-order filter is a better filter of high-frequency noise than the first-order filter.

9.3.3 Moving Average Filter

The moving average or smoothing filter realizes the average of the k last points with a fixed weight

$$y_f(n) = \frac{1}{k} \sum_{i=n-k+1}^{n} y(i) \tag{9.104}$$

giving the recursive equation

$$y_f(n) = y_f(n-1) + \frac{1}{k}[y(n) - y(n-k)] \qquad \text{with: } y_f(0) = y(0) \tag{9.105}$$

In this form, this filter would present a bias due to the initial value. For this reason, it is, in general, used with a forgetting factor ($\alpha < 1$ and close to 1)

$$y_f(n) = \alpha y_f(n-1) + \frac{1}{k}[y(n) - y(n-k)] \qquad \text{with: } y_f(0) = y(0) \tag{9.106}$$

This filter which is a lowpass filter such as the exponential filter is, in general, less efficient than the exponential filter, which gives more weight to the last measurements. The smoothing filter can be calculated by means of least squares (Mitra and Kaiser, 1993).

9.3.4 Fast Transient Filter

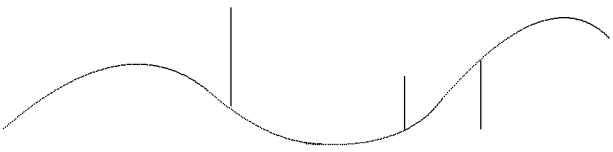


Figure 9.14: Signal presenting transient spikes

The noise can sometimes show extremely sudden transients (spikes) (Fig. 9.14), which superpose on a signal which would be smooth in their absence; in general, they are caused by the electrical environment of the sensor. They can be partially eliminated by allowing a maximum threshold Δy of the signal variation between two successive instants.

Another way to proceed is to consider the m last measurements $y(i)$, and to order these measurements by increasing amplitude. Then, the lowest m_b points and the highest m_h points are rejected; these points are later replaced in the signal by immediately lower or higher points.

Example 9.1: **Filtering of Noisy Measurements**

Some measurements in a wastewater treatment plant are obtained in a difficult environment and are noisy (Fig. 9.15). These data have been filtered with a simple filter realized as follows. First, a first-order continuous filter with unity

gain and an adequate time constant has been created and then discretized with Tustin transformation. The choice of time constant has important consequences on the aspect of the filtered data (Fig. 9.15). The time constant was chosen such that the dynamics were not too damped.

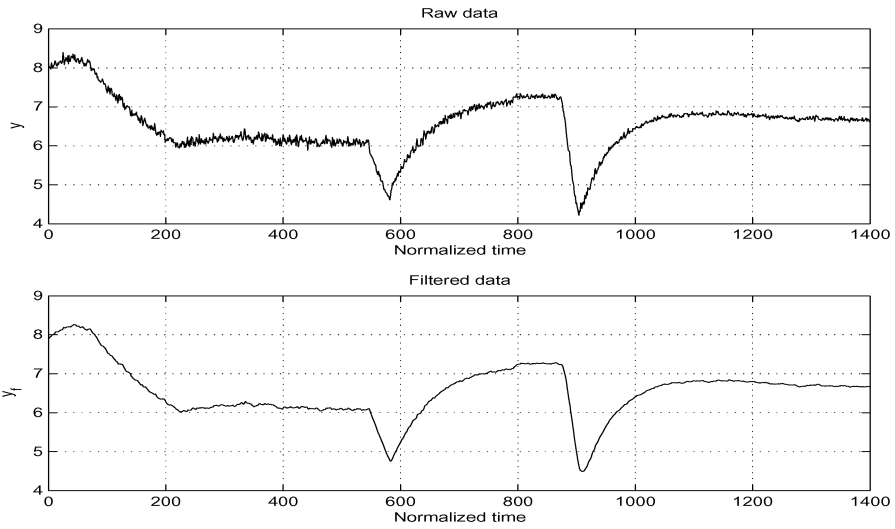


Figure 9.15: Raw data (top) from a wastewater treatment plant and filtered data (bottom)

9.4 Discrete-Time and Finite-Differences Models

By its nature, a computer cannot treat analog signals, and only sampled systems are considered. Time is discretized or discrete. Consider the continuous differential first-order equation

$$\frac{dy}{dt} = f(x, y) \tag{9.107}$$

To numerically integrate this differential equation, a large variety of schemes are possible (Carnahan et al., 1969); choose the simplest one: the explicit Euler scheme, which consists of approximating the derivative by the backward finite difference

$$\frac{y_n - y_{n-1}}{\Delta t} \approx \frac{dy}{dt} \tag{9.108}$$

The equation to be integrated becomes

$$y_n = y_{n-1} + f(x_n, y_n)\ \Delta t \tag{9.109}$$

With the implicit forward finite difference to approximate the derivative

$$\frac{y_{n+1} - y_n}{\Delta t} \approx \frac{dy}{dt} \tag{9.110}$$

a different equation would have been obtained

$$y_n = y_{n+1} - f(x_n, y_n)\ \Delta t \tag{9.111}$$

It shows that, to a unique continuous-time differential equation, correspond as many discrete differential equations as different numerical schemes exist for approximating a derivative. As this was emphasized during the statement concerning sampling, there is no bijection in the change from continuous to discrete time.

The ideal impulse (Dirac impulse) occurring at time nT_s (T_s corresponding to the sampling period) is symbolized by $\delta(t-nT_s)$. By use of simplified writing (avoiding the rigorous mathematical use of distributions), the sampled signal is described at this time as

$$y^*(nT_s) = y(nT_s)\delta(t - nT_s) \tag{9.112}$$

The sampler is qualified as an ideal impulse.

The whole signal between 0 and nT_s, sampled by a periodic train of ideal impulses, can be described by

$$\begin{array}{rcl} y^*(t) & = & y^*(0) + y^*(T_s) + \ldots + y^*(nT_s) \\ & = & y(0)\delta(t) + y(T_s)\delta(t - T_s) + \ldots + y(nT_s)\delta(t - nT_s) \\ & = & \sum_{i=0}^{n} y(iT_s)\delta(t - iT_s) \end{array} \tag{9.113}$$

The values $y(iT_s)$ correspond to constant coefficients and the Laplace transform of the sampled signal is easily deduced

$$\overline{Y}^*(s) = \sum_{i=0}^{n} y(iT_s) \exp(-iT_s s) \tag{9.114}$$

This point of view is ideal, as an impulse always has a finite length and, in this case, the Laplace transform is more complex.

Most often, a zero-order holder is used, which keeps the signal value between two sampling instants constant

$$y(t) = y(kT_s) \quad \text{if:} \quad kT_s \leq t < (k+1)T_s \tag{9.115}$$

This is referred to as the signal reconstruction. According to the type of holder used, the continuous signal thus obtained would be different.

9.5 Different Discrete Representations of a System

The Laplace transformation was a practical way for treating the systems described by continuous-time models; the z-transformation will be the method for discrete-time systems.

In order to avoid confusion, at this level we will introduce the notation $\overline{F}(s)$ for the Laplace transform of $f(t)$ so as to distinguish between the function $\overline{F}(s)$ and the z-transform denoted by $F(z)$ which does not have the same analytical form.

9.5.1 Discrete Representation: z-transform

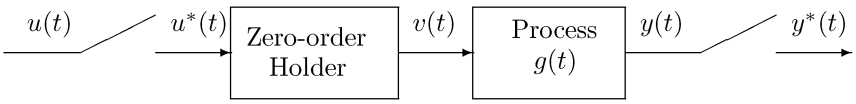


Figure 9.16: Digital control of a process

The process, being linear or linearized around its operating point, is characterized by its continuous transfer function

$$\overline{G}(s) = \frac{\overline{Y}(s)}{\overline{U}(s)} \tag{9.116}$$

To perform the digital control, before the process input, a sampler is placed so that the sampler output is $u^*(t)$ and a zero-order holder transforms the discrete signal $u^*(t)$ into a continuous signal $v(t)$ (Fig. 9.16). At the process output, the continuous signal $y(t)$ is itself sampled into $y^*(t)$ to be observed with a sampling period T_s.

Consider any sampled signal, e.g. $y^*(t)$, corresponding to the output $y(t)$ at sampling times nT_s. The sampled signal $y^*(t)$ can be represented in the form of a distribution product

$$y^*(t) = < y(t), \delta_{T_s}(t) > \tag{9.117}$$

where $\delta_{T_s}(t)$ is the periodic train of ideal impulses equal to

$$\delta_{T_s}(t) = \sum_{n=0}^{\infty} \delta(t - nT_s) \tag{9.118}$$

Thus, $y^*(t)$ can be considered as a continuous function, which is zero at every point except at sampling times nT_s where it can be represented by an impulse of amplitude $y(nT_s)$. The Laplace transform of the sampled signal $y^*(t)$ results

$$\overline{Y}^*(s) = \sum_{n=0}^{\infty} y(nT_s) \exp(-nT_s s) \tag{9.119}$$

Setting

$$z = \exp(T_s s) \tag{9.120}$$

we introduce the z-transform of the signal $y(t)$

$$\mathcal{Z}(y(t)) = Y(z) = \overline{Y}^*(s) = \sum_{n=0}^{\infty} y(nT_s) z^{-n} \tag{9.121}$$

Sometimes, the z-transform is defined as a Laurent series[3]

$$Y(z) = \sum_{n=-\infty}^{+\infty} y(nT_s) z^{-n} \tag{9.123}$$

A very important characteristic must be noted: the z-transform depends on the value of the sampling period T_s; with a given continuous signal $y(t)$, knowing T_s, we will associate a unique z-transform; on the contrary, knowing the z-transform of a signal, there exist an infinite number of continuous signals that have the same z-transform. The information contained between the sampling times is lost. The sampled output $y^*(t)$ corresponds in a unique manner to the fixed sampled input $u^*(t)$.

Different presentation (Sévely, 1969):

The sampled function $y^*(t)$ is equal to Eq. (9.117) and is simply denoted by

$$y^*(t) = y(t)\delta_{T_s}(t) \tag{9.124}$$

where $\delta_{T_s}(t)$ is the periodic train of ideal impulses.

Set $\overline{Y}(s)$, the Laplace transform of function $y(t)$. Similarly, $\overline{\Delta}_{T_s}(s)$ is the Laplace transform of function $\delta_{T_s}(t)$, which is equal to

$$\overline{\Delta}_{T_s}(s) = \mathcal{L}[\delta_{T_s}(t)] = 1 + \exp(-sT_s) + \exp(-2sT_s) + \dots \tag{9.125}$$

Provided that $|\exp(-sT_s)| < 1$, i.e. the real part of s is positive, it is possible to write

$$\overline{\Delta}_{T_s}(s) = \mathcal{L}[\delta_{T_s}(t)] = \frac{1}{1 - \exp(-sT_s)} \tag{9.126}$$

According to the convolution theorem (operator $*$), the Laplace transform of the sampled function $y^*(t)$ results

$$\begin{aligned} \overline{Y}^*(s) &= \mathcal{L}[y^*(t)] = \mathcal{L}[y(t)] * \mathcal{L}[\delta_{T_s}(t)] \quad \Longleftrightarrow \\ \overline{Y}^*(s) &= \frac{1}{2\pi j} \int_{v_R - j\infty}^{v_R + j\infty} \overline{Y}(v)\, \overline{\Delta}_{T_s}(s - v)\, dv \\ &= \frac{1}{2\pi j} \int_{v_R - j\infty}^{v_R + j\infty} \overline{Y}(v)\, \frac{1}{1 - \exp(-(s - v)T_s)}\, dv \end{aligned} \tag{9.127}$$

[3] A Laurent series is a two-sided infinite power series, e.g. mathematically

$$A(x) = \sum_{-\infty}^{+\infty} a_i x^i \tag{9.122}$$

with the condition $s_{Ry} < v_R < s_R - s_{R\delta}$ setting the notations for the complex variables:
$s = s_R + js_I$, $v = v_R + jv_I$, and
s_{Ry}: largest among the real parts (convergence abscissa of $y(t)$) of the poles of $\overline{Y}(s)$,
$s_{R\delta}$: convergence abscissa of $\delta_{T_s}(s)$, $(s_{R\delta} = 0)$.

The Laplace transform $\overline{\Delta}_{T_s}(s-v)$ of the periodic train of ideal impulses has the following poles: $p_{\delta,k} = s+j2k\pi/T_s$. On the other hand, we can assume that the poles of $\overline{Y}(s)$ have a negative real part (corresponding to a stable process)

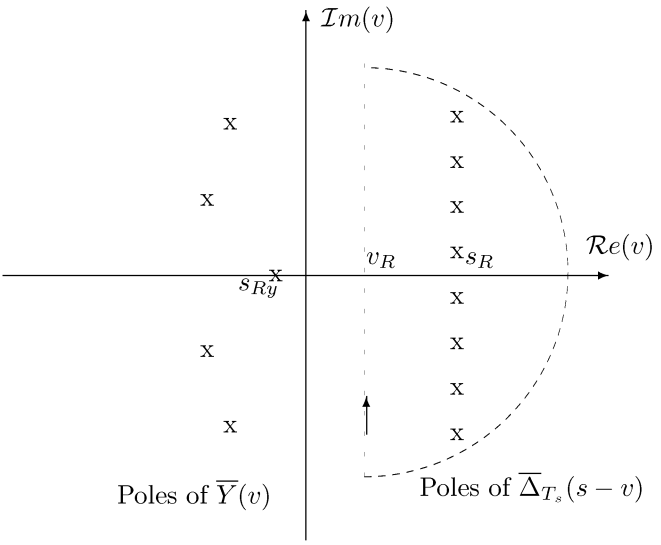


Figure 9.17: Integration by the residuals method with a contour including the poles of $\overline{\Delta}_{T_s}(s-v)$

The complex integral is calculated by the residuals method, where the integration contour (Fig. 9.17) is formed by a half-circle of infinite radius centred at $(v_R, 0)$, covered in the clockwise way. As the poles of $\overline{Y}(v)$ have a negative real part, this contour includes only the poles of $\overline{\Delta}_{T_s}(s-v)$. We obtain

$$\overline{Y}^*(s) = -\left[\sum_{p_{\delta,k}} \text{Residuals of } \overline{Y}(v)\frac{1}{1-\exp(-(s-v)T_s)}\right]_{v=p_{\delta,k}} \tag{9.128}$$

By setting the ratio of two polynomials

$$\overline{Y}(v) = \frac{N(v)}{D(v)} \tag{9.129}$$

we obtain the residual r_k relative to a given pole $p_{\delta,k}$

$$r_k = \frac{N(p_{\delta,k})}{\frac{d}{dv}[D(v)\{1-\exp(-(s-v)T_s)\}]_{v=p_{\delta,k}}} = -\frac{\overline{Y}(s+j2k\pi/T_s)}{T_s} \tag{9.130}$$

thus

$$\overline{Y^*}(s) = \frac{1}{T_s}\sum_{k=-\infty}^{\infty}\overline{Y}(s+j2k\pi/T_s) \tag{9.131}$$

assuming that $y(t) = 0$ for $t < 0$ and $y(0) = 0$. If the signal $y(t)$ is discontinuous at 0, $y(0) \neq 0$ and

$$\overline{Y^*}(s) = \frac{1}{2}y(0^+) + \frac{1}{T_s}\sum_{k=-\infty}^{\infty}\overline{Y}(s+j2k\pi/T_s) \tag{9.132}$$

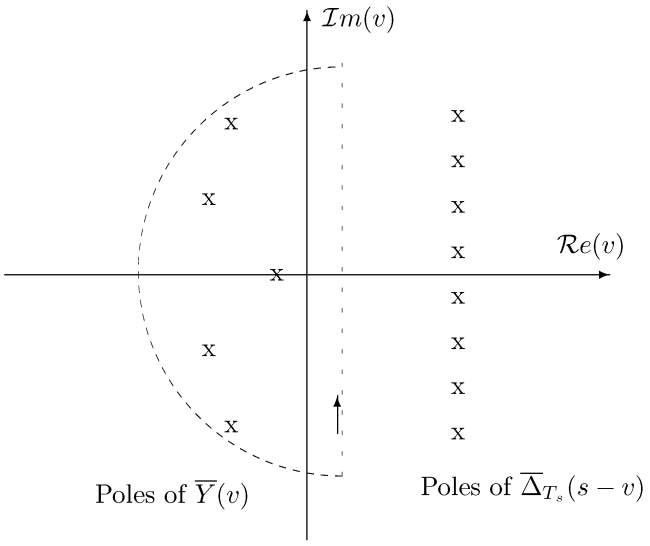


Figure 9.18: Integration by the residuals method with a contour including the poles of $\overline{Y}(v)$

Instead of taking the previous integration contour, it is possible to take as a contour a half-circle of infinite radius centred at $(v_R, 0)$, covered in the anti-clockwise way (Fig. 9.18). This contour includes the poles of $\overline{Y}(v)$, which are denoted by $p_{y,i}$, leaving aside the poles of $\overline{\Delta}_{T_s}(s-v)$. The previous expression becomes

$$\overline{Y}^*(s) = \sum_{p_{y,i}}\left[\text{ Residuals of }\ \frac{\overline{Y}(v)}{1-\exp(-(s-v)T_s}\right]_{v=p_{y,i}} \tag{9.133}$$

which can be transformed by setting $z = \exp(sT_s)$, resulting in

$$\mathcal{Z}(y(t)) = Y(z) = \sum_{p_{y,i}} \left[\text{ Residuals of } \frac{\overline{Y}(v)}{1 - \exp(vT_s)\, z^{-1}} \right]_{v=p_{y,i}} \tag{9.134}$$

This expression depends only on z and no more on s.

When the poles of $\overline{Y}(v) = \frac{N(v)}{D(v)}$ are simple, the residuals can be calculated by an expression of type (9.130), so

$$r_i = \frac{N(p_{y,i})}{\frac{d}{dv}[D(v)\,\{1 - \exp(vT_s)\, z^{-1}\}]_{v=p_{y,i}}} = \frac{N(p_{y,i})}{[D'(v)]_{p_{y,i}}\,[1 - \exp(p_{y,i}T_s)\, z^{-1}]} \tag{9.135}$$

giving the expression of the z-transform

$$\mathcal{Z}(y(t)) = Y(z) = \sum_{p_{y,i}} \frac{N(p_{y,i})}{[D'(v)]_{p_{y,i}}\,[1 - \exp(p_{y,i}T_s)\, z^{-1}]} \tag{9.136}$$

In the case where a pole $p_{y,i}$ is a multiple of order n, the residual corresponding to this pole can be calculated as

$$r_i = \frac{1}{(n-1)!} \left[\frac{d^{n-1}}{dv^{n-1}} \left((v - p_{y,i})^n \frac{\overline{Y}(v)}{1 - \exp(vT_s)\, z^{-1}} \right) \right]_{v=p_{y,i}} \tag{9.137}$$

A contour surrounding the set of the poles of $\overline{Y}(v)\,\overline{\Delta}_{T_s}(s - v)$ could have been chosen for integration.

The impulse response of the zero-order holder[4] (Fig. 9.16) is a rectangular impulse of amplitude equal to 1 and of length T_s; its transfer function is the Laplace transform of this impulse response

$$\frac{\overline{V}(s)}{\overline{U}^*(s)} = \frac{1 - \exp(-T_s\, s)}{s} \tag{9.139}$$

It results that

$$\overline{Y}(s) = \overline{G}(s) \frac{1 - \exp(-T_s\, s)}{s} \overline{U}^*(s) \tag{9.140}$$

The input $u^*(t)$ is, in reality, a sequence of impulses of amplitude $u(nT_s)$ at time nT_s

$$u^*(t) = \sum_{n=0}^{\infty} u(nT_s)\delta(t - nT_s) \tag{9.141}$$

[4]Sometimes, the zero-order holder is defined as

$$h(t) = \begin{cases} \dfrac{1}{T_s} & \text{if} \quad 0 \le t \le T_s \\ 0 & \text{otherwise} \end{cases} \tag{9.138}$$

In this way, the area is normalized to 1, in a similar manner to the continuous Dirac function as defined by physicists.

so that the output $y(t)$ is the response to this train of impulses. The response $y_n(t)$ corresponding to the input $u(nT_s)$ is such that

$$y_n(t) = u(nT_s)g(t - nT_s) \tag{9.142}$$

from the previous definition of the impulse response. The complete signal $y(t)$ is the superposition of the responses to all inputs $u(nT_s)$ so that

$$y(t) = \sum_{n=0}^{\infty} y_n(t) = \sum_{n=0}^{\infty} u(nT_s)g(t - nT_s) \tag{9.143}$$

The z-transform of the output $y(t)$ is equal to

$$Y(z) = \overline{Y}^*(s) = \sum_{n=0}^{\infty} y(nT_s)z^{-n} \tag{9.144}$$

By using Eq. (9.143) and the previous relation, the input-output relation results

$$Y(z) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} u(kT_s)g(nT_s - kT_s)z^{-n} \tag{9.145}$$

The z-transform of the input $u(t)$ is equal to

$$U(z) = \sum_{n=0}^{\infty} u(nT_s)z^{-n} \tag{9.146}$$

The relation (9.145) can be simplified: set $i = n - k$. Then, it can be separated into a product of two sequences

$$Y(z) = \sum_{i=-k}^{\infty} g(iT_s)z^{-i} \sum_{k=0}^{\infty} u(kT_s)z^{-k} \tag{9.147}$$

Taking into account the causality: the impulse response is zero for negative times: $g(iT_s) = 0$ when $i < 0$, we obtain

$$\begin{array}{rcl} Y(z) & = & \sum_{i=0}^{\infty} g(iT_s)z^{-i} \sum_{k=0}^{\infty} u(kT_s)z^{-k} \\ & = & G(z)U(z) \end{array} \tag{9.148}$$

where $G(z)$ is the z-transform of the impulse response $g(t)$ of the system consisting of the **set (zero-order holder + process)**, and is called a discrete or sampled transfer function or z-transfer function (also called impulse transfer function).

Existence Domain of z-transforms

The two-side z-transform is defined by

$$Y(z) = \sum_{n=-\infty}^{\infty} y(nT_s)z^{-n} \tag{9.149}$$

and thus appears as an extension of the one-side z-transform. This transform is useful for signals that increase exponentially, on one side only. To ensure the convergence of a power series as

$$\sum_{i=0}^{\infty} u_i \tag{9.150}$$

it is necessary and sufficient, according to Cauchy's criterion, that

$$\lim_{i\to\infty} |u_i|^{1/i} < 1 \tag{9.151}$$

This condition is applied to both series constituting the two-side z-transform

$$Y(z) = \sum_{n=-\infty}^{-1} y(nT_s)z^{-n} + \sum_{n=0}^{\infty} y(nT_s)z^{-n} = Y_-(z) + Y_+(z) \tag{9.152}$$

which gives for $Y_+(z)$

$$\lim_{n\to\infty} \left|y(nT_s)z^{-n}\right|^{1/n} < 1 \tag{9.153}$$

Supposing that

$$\lim_{n\to\infty} |y(nT_s)|^{1/n} = \rho_{y+} \tag{9.154}$$

the series $Y_+(z)$ converges when

$$|z| > \rho_{y+} \tag{9.155}$$

Changing n into $-n$, similarly the series $Y_-(z)$ converges when

$$|z| < \rho_{y-} \tag{9.156}$$

assuming that

$$\lim_{n\to\infty} |y(-nT_s)|^{-1/n} = \rho_{y-} \tag{9.157}$$

The convergence domain is thus a ring between the circles of radii ρ_{y+} and ρ_{y-} (Fig. 9.19).

z-transform of a Step Function

Consider a step $y(t)$ of amplitude A and infinite duration. The corresponding sampled function is a periodic train of ideal impulses: $y^*(t) = A\delta_{T_s}(t)$. The z-transform of $y(t)$ is equal to

$$Y(z) = \mathcal{Z}[y(t)] = A + Az^{-1} + Az^{-2} + \ldots \tag{9.158}$$

For this series to be convergent, it is necessary that

$$|z| > 1 \tag{9.159}$$

which corresponds to $s > 0$. It results that

$$\mathcal{Z}[\text{step}] = \frac{A}{1 - z^{-1}} = \frac{Az}{z - 1} \tag{9.160}$$

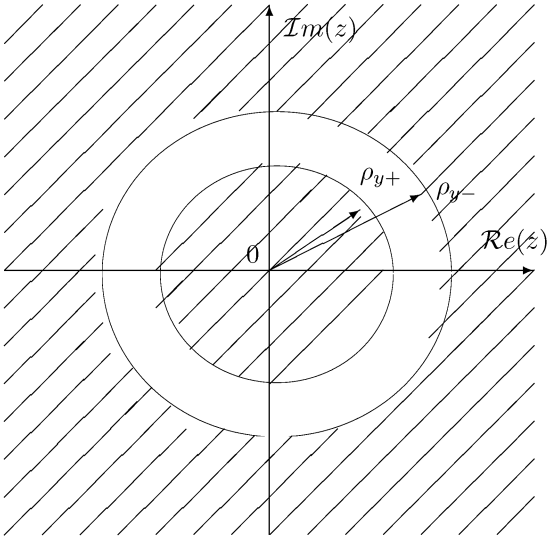


Figure 9.19: Convergence region for the two-side z-transform

z-transform of an Exponential Function

Consider the function $y(t) = A\exp(-at)$ with $a > 0$; the z-transform of this function is equal to

$$Y(z) = \sum_{n=0}^{\infty} A\exp(-anT_s)z^{-n} \tag{9.161}$$

For the series to converge, it is necessary that $|\exp(aT_s)z| > 1$, which corresponds to $s > -a$. In this case, the series converges towards

$$Y(z) = \frac{A}{1-\exp(-aT_s)z^{-1}} = \frac{Az}{z-\exp(-aT_s)} \tag{9.162}$$

By proceeding in this manner, the invariance is ensured in the time domain, i.e. the continuous function $f(t)$ and the discrete function $f(nT_s)$ perfectly coincide at the sampling times. Thus, it is possible to derive Table 9.4, which gives the z-transforms of some classical functions in parallel with their Laplace transforms.

Properties of the z-transform

Linearity:
By using the definition of the z-transform of a function, it directly results that the z-transformation is a linear transformation

$$\mathcal{Z}[a_1 f_1(t) + a_2 f_2(t)] = a_1 F_1(z) + a_2 F_2(z) \tag{9.163}$$

Translation by a real (time-delayed system):
When a process presents a time delay equal to an integer number of sampling periods: $n_r T_s$, the impulse response of this system, which can be symbolized by the three blocks: zero-order holder, process, time delay, is equal to $f(t-n_r T_s)$,

Table 9.4: z-transforms of some classical functions

Time Function	Laplace Transform	z-transform
Dirac: $\delta(t)$	1	1
Delayed Dirac: $\delta(t - nT_s)$	$\exp(-nT_s s)$	z^{-n}
Unit step: 1	$\frac{1}{s}$	$\frac{z}{z-1}$
Ramp: t	$\frac{1}{s^2}$	$\frac{T_s z}{(z-1)^2}$
Exponential: $\exp(-at)$	$\frac{1}{s+a}$	$\frac{z}{z-\exp(-aT_s)}$
Second-order	$\frac{1}{(s+a)(s+b)}$	$\frac{z(\exp(-aT_s)-\exp(-bT_s))}{(b-a)(z-\exp(-aT_s))(z-\exp(-bT_s))}$
$t\exp(-at)$	$\frac{1}{(s+a)^2}$	$\frac{T_s z\exp(-aT_s)}{(z-\exp(-aT_s))^2}$
Cosinus: $\cos(\omega t)$	$\frac{s}{s^2+\omega^2}$	$\frac{z(z-\cos(\omega T_s))}{z^2-2z\cos(\omega T_s)+1}$
Sinus: $\cos(\omega t)$	$\frac{\omega}{s^2+\omega^2}$	$\frac{z\sin(\omega T_s))}{z^2-2z\cos(\omega T_s)+1}$

with respect to the system with no time delay: zero-order holder, process, having as an impulse response $f(t)$ and a transfer function $F(s)$. The Laplace transform of the delayed system would be $F(s)\exp(-n_r T_s s)$. The z-transform of the delayed system is equal to

$$\mathcal{Z}[f(t - n_r T_s)] = z^{-n_r} F(z) \tag{9.164}$$

Translation by a real (system with advance):
Denote by n_a the integer number of sampling periods corresponding to the advance, noticing that

$$\mathcal{Z}[f(t + n_a T_s)] = \sum_{i=0}^{\infty} f[(i + n_a)T_s] z^{-i} \tag{9.165}$$

we obtain

$$\begin{aligned} \mathcal{Z}[f(t + n_a T_s)] &= z^{n_a} F(z) - z^{n_a} f(0T_s) - z^{n_a - 1} f(1T_s) - \ldots - z f((n_a - 1)T_s) \\ &= z^{n_a} \left[F(z) - \textstyle\sum_{i=0}^{n_a - 1} f(iT_s) z^{-i} \right] \end{aligned} \tag{9.166}$$

Translation by a complex number:
Consider the function $f(t)$ that contains an exponential term as a factor (this is qualified as a translation by a complex number by analogy to the similar case of the Laplace transform). The z-transform is equal to

$$\mathcal{Z}[\exp(-aT_s)f(t)] = F(z\ \exp(aT_s)) \tag{9.167}$$

as

$$\begin{array}{rcl} \mathcal{Z}[\exp(-aT_s)f(t)] & = & \sum_{n=0}^{\infty} \exp(-anT_s)f(nT_s)z^{-n} \\ & = & \sum_{n=0}^{\infty} f(nT_s)[\exp(aT_s)\ z]^{-n} \\ & = & F(z\ \exp(aT_s)) \end{array} \tag{9.168}$$

Multiplication by t^k:
It can be shown that

$$\mathcal{Z}[t^k f(t)] = -T_s z \frac{d}{dz}[F_1(z)] \quad \text{with: } F_1(z) = \mathcal{Z}[t^{k-1} f(t)] \tag{9.169}$$

which gives, for $k = 1$,

$$\mathcal{Z}[tf(t)] = -T_s z \frac{d}{dz}[F(z)] \tag{9.170}$$

Theorem of the initial value:
It is possible to know the initial value of a function from its z-transform

$$\lim_{n \to 0} f(nT_s) = \lim_{z \to \infty} F(z) \tag{9.171}$$

Theorem of the final value:
Consider

$$\mathcal{Z}[f(t + T_s) - f(t)] = zF(z) - zf(0) - F(z) \tag{9.172}$$

Use the series expansion

$$zF(z) - F(z) = (z-1) \sum_{n=0}^{\infty} f(nT_s)z^{-n} = (z-1) \lim_{n \to \infty} \sum_{k=0}^{n} f(kT_s)z^{-k} \tag{9.173}$$

On the other hand, we have

$$\begin{array}{rcl} \mathcal{Z}[f(t + T_s) - f(t)] & = & \sum_{n=0}^{\infty} [f[(n+1)T_s] - f(nT_s)]z^{-n} \\ & = & \lim_{n \to \infty} \sum_{k=0}^{n} [f[(k+1)T_s] - f(kT_s)]z^{-k} \end{array} \tag{9.174}$$

Let z tend towards 1

$$\lim_{z \to 1} \sum_{k=0}^{n} [f[(k+1)T_s] - f(kT_s)]z^{-k} = f[(n+1)T_s] - f(0) \tag{9.175}$$

It results that

$$\begin{array}{rcl} \lim_{n \to \infty}[f[(n+1)T_s] - f(0)] & = & \lim_{z \to 1} \mathcal{Z}[f(t + T_s) - f(t)] \\ & = & \lim_{z \to 1}[zF(z) - zf(0) - F(z)] \\ & = & \lim_{z \to 1}[(z-1)F(z)] - f(0) \end{array} \tag{9.176}$$

By simplification of the constant term $f(0)$, we get

$$\lim_{n\to\infty} f[(n+1)T_s] = \lim_{z\to 1}[(z-1)F(z)] \tag{9.177}$$

which constitutes the theorem of the final value, in general given in the form

$$\lim_{n\to\infty} f(nT_s) = \lim_{z\to 1}[(z-1)F(z)] \tag{9.178}$$

Theorem of summation:
The summation theorem can be written as

$$\mathcal{Z}\left[\sum_{k=0}^{n} f(kT_s)\right] = \frac{1}{1-z^{-1}}F(z) \tag{9.179}$$

Demonstration:
Consider the following series

$$g(nT_s) = \sum_{k=0}^{n} f(kT_s) \tag{9.180}$$

This series is such that

$$g(nT_s) - g[(n-1)T_s] = f(nT_s) \tag{9.181}$$

which gives the z-transform

$$\mathcal{Z}[g(nT_s)] - \mathcal{Z}[g((n-1)T_s)] = \mathcal{Z}[f(nT_s)] \tag{9.182}$$

$$G(z) - z^{-1}G(z) = F(z) \tag{9.183}$$

hence the summation theorem.
Theorem of Parseval:
This theorem is essential in signal processing, where it expresses that a signal energy is equal to the sum of the energies of its components. In the framework of the z-transform, it is expressed as

$$\sum_{n=0}^{\infty} f^2(nT_s) = \frac{1}{2\pi j}\oint_C z^{-1}F(z)F(z^{-1})dz \tag{9.184}$$

Demonstration:
As

$$f(nT_s) = \frac{1}{2\pi j}\oint_{\mathcal{C}} z^{n-1}F(z)dz \tag{9.185}$$

we can write

$$\begin{array}{rcl} \sum_{n=0}^{\infty} f^2(nT_s) & = & \sum_{n=0}^{\infty}\left[f(nT_s)\frac{1}{2\pi j}\oint_{\mathcal{C}} z^{n-1}F(z)dz\right] \\ & = & \frac{1}{2\pi j}\oint_{\mathcal{C}} F(z)\left[\sum_{n=0}^{\infty} f(nT_s)z^{n-1}dz\right] \\ & = & \frac{1}{2\pi j}\oint_{\mathcal{C}} F(z)\left[\sum_{n=0}^{\infty} f(nT_s)z^{n}\right]z^{-1}dz \\ & = & \frac{1}{2\pi j}\oint_{\mathcal{C}} F(z)F(z^{-1})z^{-1}dz \end{array} \tag{9.186}$$

Derivation and integration with respect to a parameter:

$$\mathcal{Z}\left[\frac{\partial}{\partial a}f(nT_s,a)\right]=\frac{\partial}{\partial a}F(z,a) \tag{9.187}$$

$$\mathcal{Z}\left[\int_{a_0}^{a_1}f(nT_s,a)da\right]=\int_{a_0}^{a_1}F(z,a)da \tag{9.188}$$

Limit with respect to a parameter:

$$\mathcal{Z}\left[\lim_{a\to a_0}f(nT_s,a)\right]=\lim_{a\to a_0}F(z,a) \tag{9.189}$$

Theorem of discrete convolution:

$$F_1(z)F_2(z)=\mathcal{Z}\left[\sum_{n=0}^{\infty}f_1(kT_s)f_2[(n-k)T_s]\right] \tag{9.190}$$

Inversion of z-transform

The passage from the z-transform $F(z)$ to the continuous function $f(t)$ is not unique. As a matter of fact, $F(z)$ is the z-transform of the continuous function $f^*(t)$ obtained by sampling with period T_s and the zero-order holder of the function $f(t)$. The information loss between two sampling instants forbids the reconstruction of the function $f(t)$. The inverse transformation will be denoted by $\mathcal{Z}^{-1}$

$$f^*(t)=\{f(nT_s)\}=\mathcal{Z}^{-1}[F(z)] \tag{9.191}$$

Several methods allow us to obtain the set $\{f(nT_s)\}$, such as the residuals method, polynomial division, expansion as a sum of rational fractions or simply numerical calculation on a computer.

a) Method of residuals. This method relies on a theorem by Cauchy

$$\frac{1}{2\pi j}\oint_{\mathcal{C}}z^k dz=\begin{cases}1 & \text{for } k=-1\\ 0 & \text{for } k\neq -1\end{cases} \tag{9.192}$$

$\mathcal{C}$ being a contour surrounding the z-plane origin.

On the other hand, use the expansion of $F(z)$

$$F(z)=\sum_{n=0}^{\infty}f(nT_s)z^{-n} \tag{9.193}$$

To display the term $f(nT_s)$, multiply on both sides by z^{n-1}, and integrate over $\mathcal{C}$, which gives

$$f(nT_s)=\frac{1}{2\pi j}\oint_{\mathcal{C}}z^{n-1}F(z)dz \tag{9.194}$$

This integral can be calculated according to the residuals method; choose a contour surrounding the z-plane origin and the poles p_i of $F(z)$

$$f(nT_s)=\sum_{p_i}\left[\text{ residuals of } z^{n-1}F(z)\right]_{z=p_i} \tag{9.195}$$

b) Polynomial division. The function $F(z)$ is often presented as a rational fraction with respect to z

$$F(z) \quad = \frac{B(z)}{A(z)} = \frac{b_0\, z^{n_b} + b_1\, z^{n_b-1} + \ldots + b_{n_b}}{a_0\, z^{n_a} + a_1\, z^{n_a-1} + \ldots + a_{n_a}} \tag{9.196}$$

with the condition $n_a \geq n_b$ so that the transfer function is physically realizable: the controlled output must not precede the manipulated input. For example,

$$F(z) = \frac{z^2 - 0.5}{z - 1}$$

is not physically realizable.

The function $F(z)$ is still more often expressed as a rational fraction with respect to z^{-1}; this ratio of two polynomials can be expanded as a series in z^{-1}. Thus

$$\begin{aligned} F(z) \quad &= \frac{B(z^{-1})}{A(z^{-1})} = \frac{b_0 + b_1\, z^{-1} + \ldots + b_{n_b}\, z^{-n_b}}{a_0 + a_1\, z^{-1} + \ldots + a_{n_a}\, z^{-n_a}} \\ &= f(0) + f(T_s)\, z^{-1} + \ldots + f(nT_s)\, z^{-n} + \ldots \end{aligned} \tag{9.197}$$

with

$$f(kT_s) = \frac{1}{a_0}\,[b_k - \sum_{i=0}^{k-1} a_{k-i} f(iT_s)] \quad \text{with } b_k = 0 \text{ for } k > n_b \tag{9.198}$$

Example 9.2: **Polynomial Division**

Given the following discrete transfer function

$$F(z) = \frac{2\,z - 3}{3\,z^2 + 2\,z - 1} \tag{9.199}$$

find the first values of $f_k = f(kT_s)$ by polynomial division.

The polynomial division is represented by

$$\begin{array}{llll|lllll} 2z & -3 & & & 3z^2 & +2z & -1 & & \\ \hline -2z & -\frac{4}{3} & +\frac{2}{3}z^{-1} & & 0 & +\frac{2}{3}z^{-1} & -\frac{13}{9}z^{-2} & +\frac{32}{27}z^{-3} & \\ \cline{1-3} & -\frac{13}{3} & +\frac{2}{3}z^{-1} & & & & & & \\ & +\frac{13}{3} & +\frac{26}{9}z^{-1} & -\frac{13}{9}z^{-2} & & & & & \\ \cline{2-4} & & +\frac{32}{9}z^{-1} & -\frac{13}{9}z^{-2} & & & & & \end{array}$$

It results that the first values of $f(kT_s)$ are

$$f_0 = 0 \quad ; \quad f_1 = \frac{2}{3} \quad ; \quad f_2 = -\frac{13}{9} \quad ; \quad f_3 = \frac{32}{27}$$

c) Sum of rational fractions. In an analogous manner to that used for the Laplace transformation, knowing that $F(z)$ possesses n_a poles, $F(z)$ is expanded as a sum of rational fractions with respect to z^{-1} according to

$$F(z) = \frac{c_1}{1 - p_1\, z^{-1}} + \ldots + \frac{c_{n_a}}{1 - p_{n_a}\, z^{-1}} \tag{9.200}$$

thus

$$f(nT_s) = \mathcal{Z}^{-1}\left[\frac{c_1}{1 - p_1\, z^{-1}}\right] + \ldots + \mathcal{Z}^{-1}\left[\frac{c_{n_a}}{1 - p_{n_a}\, z^{-1}}\right] \tag{9.201}$$

and

$$f(nT_s) = c_1 p_1^{nT_s} + \ldots + c_{n_a} p_{n_a}^{nT_s} \tag{9.202}$$

The scalar coefficients c_i are real or conjugate complex.

Relation between Discrete Transmittance and Difference Equations

A differential equation can be discretized by any finite-difference method (Euler, forward or backward differences, etc.). The system is then represented by the general equation with constant coefficients a_i and b_i

$$a_0 y_n + a_1 y_{n-1} + \ldots + a_{n_a} y_{n-n_a} = b_0 u_n + b_1 u_{n-1} + \ldots + b_{n_b} u_{n-n_b} \tag{9.203}$$

where the initial conditions should be specified. Using the fact that

$$\mathcal{Z}(y_{n-i}) = z^{-i} Y(z) \tag{9.204}$$

we pass in a bijective, thus unique, manner from the finite-difference equation to the discrete transfer function

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_0 + b_1 z^{-1} + \ldots + b_{n_b} z^{-n_b}}{a_0 + a_1 z^{-1} + \ldots + a_{n_a} z^{-n_a}} \tag{9.205}$$

In general, as the output y is not immediately influenced by the input, we get $b_0 = 0$. The concept of physical realizability is expressed by the fact that the output depends only on the past inputs: thus, y_n cannot depend on u_{n+1}.

Example 9.3: **Discrete PID Controller**

The analog PID controller corresponds to the following equation between the deviation variables ($\tilde{e}(t)$ controller input and $\tilde{r}(t)$ controller output)

$$\tilde{r}(t) = K_r \left[\tilde{e}(t) + \frac{1}{\tau_I}\int_0^t \tilde{e}(x)dx + \tau_D \frac{d\tilde{e}(t)}{dt}\right] \tag{9.206}$$

giving, for example, the following difference equation (here based on a backward and non-unique difference)

$$r_n - r_{n-1} = K_r \left[e_n - e_{n-1} + \frac{T_s}{\tau_I} \frac{(e_n + e_{n-1})}{2} + \frac{\tau_D}{T_s}(e_n - 2e_{n-1} + e_{n-2}) \right] \tag{9.207}$$

which gives the discrete transfer function

$$G(z) = \frac{R(z)}{E(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 - z^{-1}} \tag{9.208}$$

with $b_0 = K_r(1 + \frac{T_s}{2\tau_I} + \frac{\tau_D}{T_s})$, $b_1 = K_r(-1 + \frac{T_s}{2\tau_I} - \frac{2\tau_D}{T_s})$, $b_2 = K_r \frac{\tau_D}{T_s}$.

It would have been possible to consider the continuous transfer function associated with the PID

$$\overline{G}(s) = K_r \left[1 + \frac{1}{\tau_I s} + \tau_D s \right] \tag{9.209}$$

By applying the z-transform according to Table 9.4, a different expression of the discrete transfer function results

$$G(z) = K_r \left[1 + \frac{z}{\tau_I (z-1)} + \tau_D \frac{z-1}{z} \right] \tag{9.210}$$

Stability Analysis

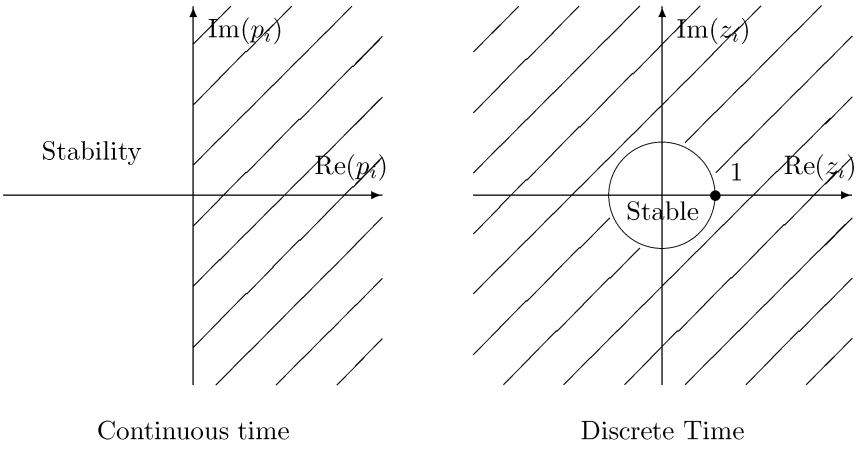


Figure 9.20: Stability domains (position of the poles)

Consider the system described by the discrete transfer function (with $n_a \geq n_b$ for the z-transfer function)

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_0 + b_1 z^{-1} + \ldots + b_{n_b} z^{-n_b}}{a_0 + a_1 z^{-1} + \ldots + a_{n_a} z^{-n_a}} = \frac{N_g(z)}{D_g(z)} \tag{9.211}$$

Submit this system to a stable input $u(t)$ defined by its z-transform as a ratio of two polynomials

$$U(z) = \frac{N_u(z)}{D_u(z)} \qquad , \quad \text{with: } \deg(N_u) \leq \deg(D_u) \tag{9.212}$$

The z-transform of the output is equal to

$$\begin{aligned} Y(z) &= G(z)\, U(z) = \frac{N_g(z)}{D_g(z)} \frac{N_u(z)}{D_u(z)} \\ &= \frac{N_1(z)}{D_g(z)} + \frac{N_2(z)}{D_u(z)} \\ &= Y_n(z) + Y_f(z) \end{aligned} \tag{9.213}$$

In the expression of the transform of the output $Y(z)$, in the transform $Y_f(z)$ of the forced response we recognize the influence of the input $u(t)$ through the stable denominator $D_u(z)$, thus the forced response $y_f(t)$ is stable. There remains the transform $Y_n(z)$ of the natural response whose stability is conditioned by the denominator $D_g(z)$ of the discrete transfer function $G(z)$.

The transform $Y_n(z)$ of the natural response can be expanded as a sum of rational fractions

$$Y_n(z) = \sum_{i=1}^{n_a} \frac{c_i}{z - z_i} \tag{9.214}$$

where z_i are the roots of the denominator of $G(z)$, thus the poles of $G(z)$. The poles of $G(z)$ are the z values for which the discrete transfer function $G(z)$ tends towards infinity, the zeros of $G(z)$ being the z values for which the discrete transfer function $G(z)$ tends towards 0. Consider the case where only one pole exists, so that

$$Y_n(z) = \frac{c_1}{z - z_1} = \sum_{n=0}^{\infty} y(nT_s) z^{-n} \tag{9.215}$$

The rational fraction can be expanded as a series, which converges only if $|z_1/z| < 1$

$$Y_n(z) = \frac{c_1}{z - z_1} = z^{-1} \frac{c_1}{1 - \dfrac{z_1}{z}} = z^{-1} c_1 [1 + (\frac{z_1}{z}) + (\frac{z_1}{z})^2 + \ldots] \tag{9.216}$$

hence

$$y_n(nT_s) = c_1 (z_1)^{n-1} \tag{9.217}$$

Thus, the condition of stability is that the z poles of the discrete transfer function are situated inside the circle of unit radius in the complex plane (Figs. 9.20 and 9.21).

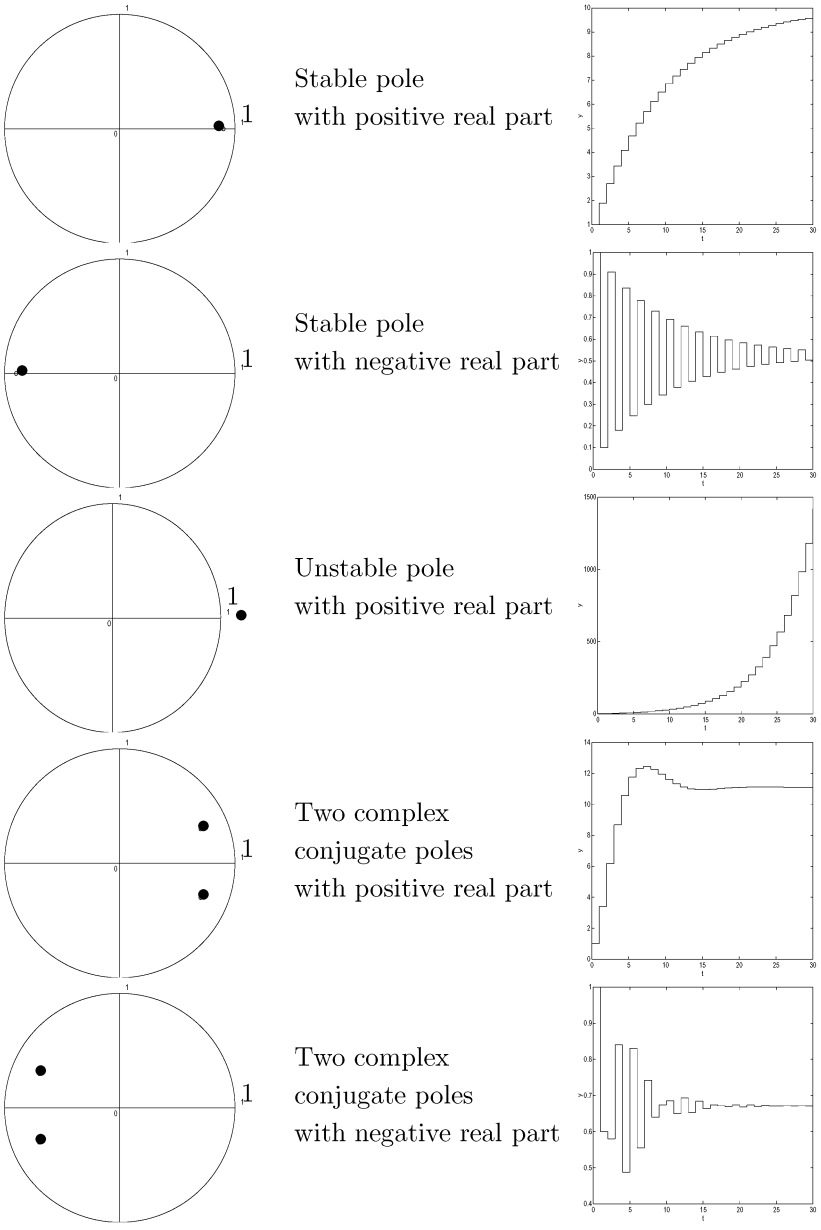


Figure 9.21: Influence of the position of the poles of a discrete transfer function on a step response

The analogy of this condition for discrete time with the parallel condition for continuous time can be realized: in fact, the continuous-time stability demands that the poles are located in the left complex half-plane corresponding to $\mathcal{R}e(s) < 0$. The transformation

$$z_i = \exp(T_s s_i) \tag{9.218}$$

then gives the discrete-time condition

$$|z_i| < 1 \tag{9.219}$$

In practice, stability is, of course, essential, but performance and robustness must also be taken into account. For this reason, in the same manner as in continuous time, the requirements of gain and phase margins are such that all of the left half-plane is not favourable (avoid being too close to the imaginary axis); in discrete time, only one part of the unit circle is recommended. The influence of the position of the poles on the response to a step input appears clearly in Fig. 9.21.

The condition of stability of the transfer function $G(q)$ is stated as

$$G(q) = \sum_{k=1}^{\infty} g(k) q^{-k} \text{ is stable } \iff \sum_{k=1}^{\infty} |g(k)| < \infty \tag{9.220}$$

The consequence is that the associated series expansion

$$G(z) = \sum_{k=1}^{\infty} g(k) z^{-k} \text{ converges } \forall\ |z| \geq 1 \tag{9.221}$$

This function $G(z)$ is said to be analytic on the unit circle and outside the unit circle: $G(z)$ has no poles on the circle and outside the circle.

Criterion of stability of Jury. The criterion of stability of Jury provides necessary and sufficient conditions for a real-coefficient polynomial to have its roots inside the unit circle.

Consider the polynomial

$$P(z) = a_0 z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n \tag{9.222}$$

whose coefficient a_0 is positive, then compile the associated Table 9.5.

The two first rows of Table 9.5 simply contain the coefficients of the polynomial taken in the opposite order. The following row of the table (coefficients b_i) is obtained by subtracting from the first row the second row multiplied by the ratio r_1, and so on until the row $2n + 1$.

The necessary and sufficient conditions for the roots of $P(z)$ (where $a_0 > 0$ is imposed) to be situated inside the unit circle are that the head coefficients a_0, b_0, c_0, ..., are positive.

Table 9.5: Table of the stability criterion of Jury

Row								
1	a_0	a_1	a_2	$\dots$	a_{n-2}	a_{n-1}	a_n	
2	a_n	a_{n-1}	a_{n-2}	$\dots$	a_2	a_1	a_0	$r_1 = \dfrac{a_n}{a_0}$
3	b_0	b_1	b_2	$\dots$	b_{n-2}	b_{n-1}	0	(1st row of a_i) $- r_1$ (2nd row of a_i)
4	b_{n-1}	b_{n-2}	b_{n-3}	$\dots$	b_1	b_0		$r_2 = \dfrac{b_{n-1}}{b_0}$
5	c_0	c_1	c_2	$\dots$	c_{n-2}	0		(1st row of b_i) $- r_2$ (2nd row of b_i)
6	c_{n-2}	c_{n-3}	c_{n-4}	$\dots$	c_0			$r_3 = \dfrac{c_{n-2}}{c_0}$
$\vdots$								
$2n+1$	w_0	0						

Example 9.4: **Stability of a Second-Order Transfer Function**

A classical example concerns the stability of the sampled following second-order transfer function

$$G(z) = \frac{b_0 z^2 + b_1 z + b_2}{z^2 + a_1 z + a_2} \tag{9.223}$$

Table 9.6: Table of Jury stability criterion in the case of a discrete second-order transfer function

Row				
1	1	a_1	a_2	
2	a_2	a_1	1	$r_1 = \dfrac{a_2}{1}$
3	$1 - a_2^2$	$a_1 - a_1 a_2$	0	
4	$a_1 - a_1 a_2$	$1 - a_2^2$		$r_2 = \dfrac{a_1 - a_1 a_2}{1 - a_2^2} = \dfrac{a_1}{1 + a_2}$
5	$1 - a_2^2 - (a_1 - a_1 a_2)\dfrac{a_1}{1 + a_2}$	0		

Jury Table 9.6 results yield the following two conditions

$$\begin{aligned} 1 - a_2^2 &> 0 \iff |a_2| < 1 \\ 1 - a_2^2 - (a_1 - a_1 a_2)\frac{a_1}{1 + a_2} &> 0 \iff (1 + a_2)^2 - a_1^2 > 0 \iff 1 + a_2 > |a_1| \end{aligned} \tag{9.224}$$

The stability domain with respect to both parameters a_1 and a_2 resulting from these conditions is the inside of the triangle of Fig. 9.22.

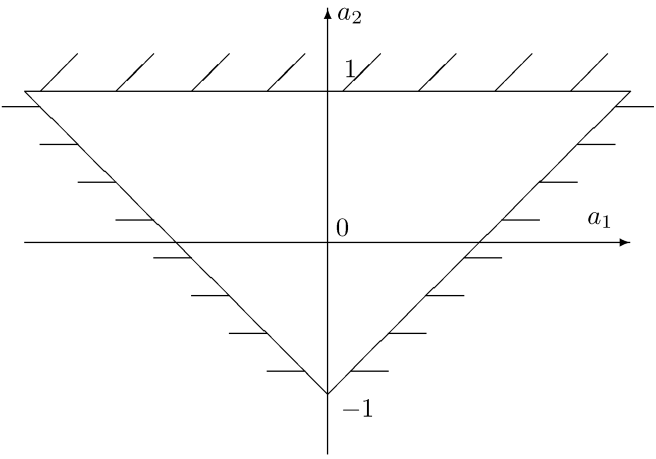


Figure 9.22: Stability domain of the discrete second-order transfer function with respect to its parameters

---

Causality and time delay. Consider a discrete transfer function

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_0 + b_1 z + \ldots + b_{n_b} z^{n_b}}{a_0 + a_1 z + \ldots + a_{n_a} z^{n_a}} \tag{9.225}$$

A system is causal when the output depends only on the past inputs. This physical condition implies for the discrete transfer function, expressed as a ratio of two polynomials with respect to z (not to z^{-1} !), that

$$n_a \geq n_b \tag{9.226}$$

Such a discrete transfer function is proper.

The relative degree r is defined by

$$r = n_a - n_b \tag{9.227}$$

If the relative degree r is zero, the discrete transfer function is biproper and the output reacts without delay to an input variation. If the relative degree is strictly positive, the discrete transfer function is strictly proper and the output reacts with a delay of r sampling periods to an input variation.

9.5.2 Conversion of a Continuous Description in Discrete Time

Suppose that a process is known by its continuous transfer function $\overline{G}(s)$. It is interesting to deduce a corresponding discrete transfer function. In fact, as the discretization of a continuous differential equation is not unique, the discrete representation depends on the mode of transformation chosen. A given correspondence between s and z will be used allowing us to transform a continuous transfer function into a discrete transfer function and the transformation will be qualified as a frequency one.

Here, let us cite several possible transformations which allow us to get a discrete transfer function $G(z)$, knowing the continuous transfer function $\overline{G}(s)$. T_s represents the sampling period.

Consider a continuous system described by the state equation

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{9.228}$$

subjected to a zero input. Integrate the resulting differential equation between two instants t_0 and t_1. Moreover, assume that $t_1 = t_0 + T_s$. If we approximate the integral of $Ax(t)$ by the area of the rectangle of length T_s and height $x(t_0)$, we obtain the approximation

$$x(t_0 + T_s) - x(t_0) \approx Ax(t_0)T_s \Longleftrightarrow \frac{x(t_0 + T_s) - x(t_0)}{T_s} \approx Ax(t_0) \tag{9.229}$$

hence

$$\dot{x}(t) \approx \frac{x(t + T_s) - x(t)}{T_s} \tag{9.230}$$

leading to the following formula of forward difference.
Forward difference (integration) or implicit Euler method:

$$s = \frac{z - 1}{T_s} \tag{9.231}$$

The stability of $\overline{G}(s)$ is not maintained by this transformation, which is equivalent to placing a zero-order holder at the input.
Backward difference (integration) or explicit Euler method:

$$s = \frac{z - 1}{T_s z} \tag{9.232}$$

The stability of $\overline{G}(s)$ is maintained by this transformation.
Bilinear or trapezoidal integration, or Tustin method:

$$s = \frac{2}{T_s}\frac{z - 1}{z + 1} \Longleftrightarrow z = \frac{1 + \dfrac{T_s}{2}s}{1 - \dfrac{T_s}{2}s} \tag{9.233}$$

This mapping: $s \rightarrow z$ is bijective (to any value of s corresponds one and only one z, and vice versa). It offers the advantage of making the imaginary axis of the complex plane of s (stability limit in continuous time) correspond to the unit circle of the complex plane of z (stability limit in discrete time). Moreover, to the left half-plane of s (stability region in continuous time) corresponds the inside of the unit circle for z (stability region in discrete time). The Tustin method is frequently used. In particular, it allows us to perform the frequency analysis of a discrete-time system from a continuous-time system by the transformation of s into $z = \exp(j\omega T_s)$, where ω is the frequency of the digital system associated with the frequency $\overline{\omega}$ of the analog system by the relation

$$\overline{\omega} = \frac{2}{T_s} \tan(\frac{\omega T_s}{2}) \tag{9.234}$$

To the domain $[0, +\infty[$ of the analog frequency $\overline{\omega}$ corresponds the domain $[0, +\pi/T_s[$ of the digital frequency ω (Fig. 9.23); the figure might have been extended to the domain of negative frequencies which is symmetric with respect to the axis origin. This compression from analog to digital frequency is called frequency warping. While the unit of analog frequency is rad/s, the unit of digital frequency is rad/sampling interval. Thus, the highest digital frequency is π rad/sampling interval and is called the Nyquist frequency. In different units, it is equal to

$$\begin{aligned} \text{Nyquist frequency} &= \pi \text{ rad/sampling interval} \\ &= \frac{\pi}{T_s} \text{ rad/s} \\ &= \frac{1}{2T_s} \text{ Hz} \\ &= \frac{1}{2} \text{ revolutions/s} \end{aligned} \tag{9.235}$$

The normalized frequency (which is not always normalized in the same manner!) takes as reference the Nyquist frequency, where the latter can be indicated as being equal to π (refer to rad/sampling interval units) or 0.5 (refer to revolutions/s units).

In a more general formulation (Mitra and Kaiser, 1993), the Tustin transformation can be expressed as

$$s = \alpha \frac{z-1}{z+1} \Longleftrightarrow z = \frac{1 + \frac{s}{\alpha}}{1 - \frac{s}{\alpha}} \tag{9.236}$$

where α would take the value $\alpha = 2/T_s$ in the previous Tustin transformation. The stability properties mentioned above are maintained.

Method of pole-zero correspondence:

Consider a continuous transfer function $\overline{G}(s)$ having poles p_i and zeros q_i. According to the formula $z = \exp(sT_s)$, each pole p_i is transformed into $\exp(p_i T_s)$ and each zero q_i into $\exp(q_i T_s)$. Moreover, the steady-state gains of the continuous and discrete transfer functions are equal.

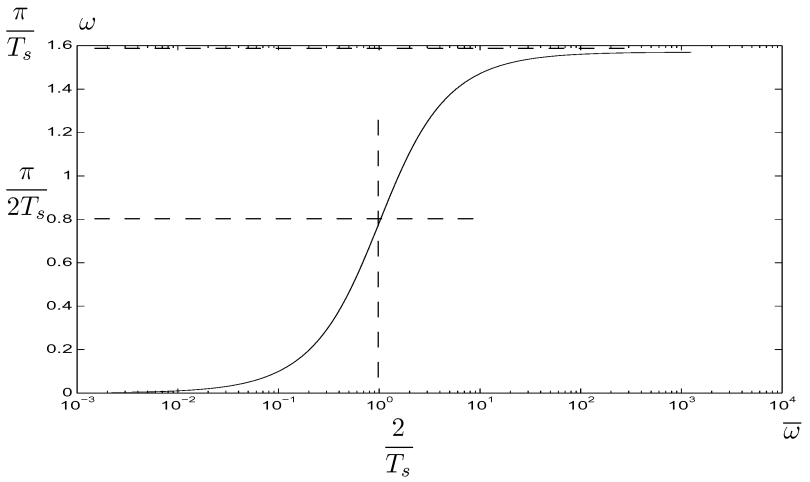


Figure 9.23: Correspondence between the analog frequency $\overline{\omega}$ and the digital frequency ω for $T_s = 2$

Example 9.5: **Transformation by Pole-Zero Correspondence**

By using the method of pole-zero correspondence, to the following continuous transfer function

$$\overline{G}(s) = \frac{5(s-1)}{(s+2)(s+3)} \tag{9.237}$$

corresponds the discrete transfer function

$$G(z) = \frac{5(z - \exp(T_s))}{(z - \exp(-2T_s))(z - \exp(-3T_s))} \tag{9.238}$$

In this form, the gain of the discrete transfer function would be different from the steady-state gain of the continuous transfer function. It suffices to multiply the numerator of the discrete transfer function by a coefficient c to ensure the same steady-state gain.

The drawback of the pole-zero correspondence is that if the continuous transfer function possesses more poles than zeros, the discrete transfer function will have a relative degree equal to the difference of the number of poles and zeros, so that it introduces a time delay equal to the relative degree. For this reason, often a polynomial of degree $r-1$ is artificially introduced in the numerator in order to bring back the delay to the minimum delay of one sampling period of the output with respect to the input.

In Table 9.7, some discrete transfer functions corresponding to continuous transfer functions are given in the case of backward difference (zero-order holder). Differences may be noted by comparison with classical references. Table 9.7 is in agreement with numerical results given by Matlab©, in particular with respect to the steady-state gain.

Table 9.7: Continuous-discrete transformation by assuming a zero-order holder at the input (backward difference of Euler type)

Time Function $g(t)$	Continuous Transfer Function $\overline{G}(s)$	Discrete Transfer Function G(z)
Dirac: $\delta(t)$	1	1
Delayed Dirac: $\delta(t - nT_s)$	$\exp(-nT_s s)$	z^{-n}
Unit step: 1	$\frac{1}{s}$	$\frac{zT_s}{z-1}$
Ramp: t	$\frac{1}{s^2}$	$\frac{T_s^2}{2}\frac{z+1}{(z-1)^2}$
Exponential: $\exp(-at)$	$\frac{1}{s+a}$	$\frac{c}{z-\exp(-aT_s)}$
Second-order: $\exp(-at)$	$\frac{1}{(s+a)(s+b)}$	$\frac{c(z+1)}{(z-\exp(-aT_s))(z-\exp(-bT_s))}$

9.5.3 Operators

Operator q:
Instead of z-transformation, the forward shift operator q or the backward shift operator q^{-1} can be defined as

$$y(t+1) = q\,y(t) \qquad \text{or:} \qquad y(t-1) = q^{-1}y(t) \tag{9.239}$$

The difference equation of a linear system

$$a_0 y_n + a_1 y_{n-1} + \ldots + a_{n_a} y_{n-n_a} = b_0 u_n + b_1 u_{n-1} + \ldots + b_{n_b} u_{n-n_b} \tag{9.240}$$

thus becomes

$$\left[a_0 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a}\right] y(t) = \left[b_0 + b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b}\right] u(t) \tag{9.241}$$

which will be simply denoted by

$$A(q)y(t) = B(q)u(t) \tag{9.242}$$

where $A(q)$ and $B(q)$ are the following polynomials, in fact, depending on q^{-1} in the form

$$\begin{array}{rcl} A(q) & = & a_0 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a} \\ B(q) & = & b_0 + b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b} \end{array} \tag{9.243}$$

The ratio of both polynomials will be considered as the discrete transfer operator of the discrete transfer function (strictly speaking, $G(z)$ should be used in the latter case) of the system. For linear systems, the forward shift operator q and the variable z defining the z-transform are equivalent, but the operator

q is defined only by Eq. (9.239), so that it can be applied to any discrete-time system, thus as well to nonlinear systems.

The variable z is analytical: we can speak of numerical values z_i of the poles of a transfer function $G(z)$. The operator q does not possess any numerical value; it gives the transfer function $G(q)$, whose mathematical expression is strictly identical to $G(z)$.

Operator δ:

The operator δ is defined from the operator q by the relation

$$\delta = \frac{q-1}{T_s} \tag{9.244}$$

where T_s is the sampling period. With the relation between q and δ being linear, all operations previously realized with the operator q (or z for a linear system) can be performed with operator δ. In particular, this definition can be used to transform a discrete transfer function $G(q)$ (or $G(z)$ in this linear case) into a transfer function $H(\delta)$. The stability condition $z < 1$ becomes $1 + \delta T_s < 1$. In state space, by means of the operator δ, a linear system can be written in the general form analogous to Eq. (7.1)

$$\left\{ \begin{array}{ll} \delta \boldsymbol{x}_k & = \dfrac{\boldsymbol{x}_{k+1} - \boldsymbol{x}_k}{T_s} = \boldsymbol{A}_\Delta \boldsymbol{x}_k + \boldsymbol{B}_\Delta u_k \\ y_k & = \boldsymbol{C}_\Delta \boldsymbol{x}_k \end{array} \right. \tag{9.245}$$

This system is equivalent to the discrete system

$$\left\{ \begin{array}{ll} \boldsymbol{x}_{k+1} & = (\boldsymbol{I} + T_s \boldsymbol{A}_\Delta)\boldsymbol{x}_k + T_s \boldsymbol{B}_\Delta u_k \\ y_k & = \boldsymbol{C}_\Delta \boldsymbol{x}_k \end{array} \right. \tag{9.246}$$

With regard to the matrices of the continuous system, the matrices $\boldsymbol{A}_\Delta$ and $\boldsymbol{B}_\Delta$ are respectively equal to

$$\boldsymbol{A}_\Delta = \frac{\exp(\boldsymbol{A}T_s) - \boldsymbol{I}}{T_s} \; ; \quad \boldsymbol{B}_\Delta = \frac{\int_0^{T_s} \exp(\boldsymbol{A}t)\boldsymbol{B}dt}{T_s} \tag{9.247}$$

Example 9.6: **Discretization of a Continuous Second-Order Transfer Function**

Consider, for example, a continuous second-order transfer function

$$\overline{G}(s) = \frac{1}{9s^2 + 3s + 1} \tag{9.248}$$

Two types of transformation have been applied to this transfer function to obtain a discrete transfer function: sampling with a zero-order holder and a Tustin bilinear transformation (calculations with Matlab©). The left column of Table 9.8 results. In both cases, different sampling periods T_s have been used. Then, the operator δ has been applied to each of the discrete transfer functions. It appears that, unlike other z-transfer functions, the denominator

Table 9.8: Transformations from a continuous transfer function into a discrete transfer function by two different transformations and resulting δ-transforms

Sampling Period	Zero-Order Holder	δ-Transform
$T_s = 1$	$\frac{0.0494z + 0.0442}{z^2 - 1.6229z + 0.7165}$	$\frac{0.4447\delta + 0.8426}{9\delta^2 + 3.3938\delta + 0.8426}$
$T_s = 0.1$	$\frac{10^{-4}(5.494z + 5.433)}{z^2 - 1.9661z + 0.9672}$	$\frac{0.0494\delta + 0.9834}{9\delta^2 + 3.0489\delta + 0.9834}$
$T_s = 0.01$	Period T_s too low	$\frac{0.0050\delta + 0.9983}{9\delta^2 + 3.0050\delta + 0.9983}$
Sampling Period	Bilinear Transformation (Tustin)	δ-Transform
$T_s = 1$	$\frac{0.0233z^2 + 0.0465z + 0.0233}{z^2 - 1.6279z + 0.7209}$	$\frac{0.2093}{9\delta^2 + 3.3488\delta + 0.837}$
$T_s = 0.1$	$\frac{10^{-4}(2.731z^2 + 5.463z + 2.731)}{z^2 - 1.9661z + 0.9672}$	$\frac{0.2458}{9\delta^2 + 3.0483\delta + 0.9833}$
$T_s = 0.01$	Period T_s too low	$\frac{0.2496}{9\delta^2 + 3.0050\delta + 0.9983}$

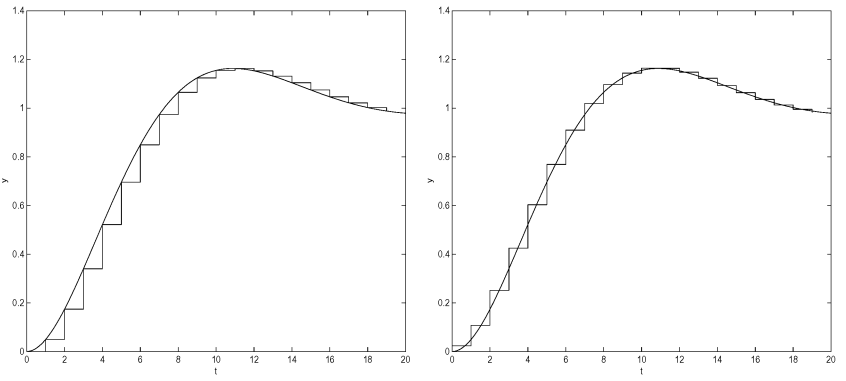


Figure 9.24: Comparison of the step responses resulting from a continuous transfer function and from the discrete transfer function with a zero-order holder (left) or from the discrete transfer function with Tustin transformation (right)

of δ transfer functions is close to that of the continuous transfer function and gets all the closer because the sampling period is low.

In Fig. 9.24, the step responses obtained from the continuous transfer func-

tion, from the discrete transfer function with a zero-order holder (response A) with sampling period $T_s = 1$, and from the discrete transfer function with Tustin transformation (response B), are compared. It appears clearly that the step response A remains under the continuous curve while the step response B is shared on both sides of the continuous curve.

In Example 9.6, we thus empirically find again the fact that the operator δ tends towards the derivation operator when the sampling period tends towards zero

$$\lim_{T_s \to 0} \delta = \frac{d}{dt} \tag{9.249}$$

Similarly, if the models were expressed in state-space form, the coefficients of the matrices of the δ model and of the continuous s model would be close.

To simulate a known discrete system as a discrete z-transfer function, De Larminat (1993) recommends that in order to improve the numerical robustness to go through the intermediary of the δ model according to:

1. Conversion of the discrete z-transfer function into a discrete δ-transfer function according to $H(\delta) = G(1 + \delta T_s)$.
2. Passage from the discrete δ transfer function to the discrete δ state-space model and simulation of this model (Middleton and Goodwin, 1986).

A reference textbook on the interest in δ operator is by Middleton and Goodwin (1990). In particular, the link continuous-discrete is emphasized at high sampling frequencies, with the improvement of the numerical properties by comparison with the operator q. The properties of the δ-transform as well as the δ-transforms for a certain number of usual functions are given.

References

H. Baher. *Analog and Digital Signal Processing.* Wiley, 2nd edition, 2000.

M. Bellanger. *Analyse des Signaux et Filtrage Numérique Adaptatif.* Masson, Paris, 1989.

E.O. Brigham. *The Fast Fourier Transform.* Prentice Hall, Englewood Cliffs, New Jersey, 1974.

B. Carnahan, H.A. Luther, and J.O. Wilkes. *Applied Numerical Methods.* Wiley, New York, 1969.

C.T. Chen. *One-dimensional Digital Signal Processing.* Marcel Dekker, New York, 1979.

R.E. Crochiere and L.R. Rabiner. *Multirate Digital Signal Processing.* Prentice Hall, 1983.

P. De Larminat. *Automatique, Commande des Systèmes Linéaires.* Hermès, Paris, 1993.

J.M. Flaus. *La Régulation Industrielle.* Hermès, Paris, 1994.

G.C. Goodwin and K.S. Sin. *Adaptive Filtering, Prediction and Control.* Prentice Hall, Englewood Cliffs, 1984.

S. Haykin. *Adaptive Filter Theory.* Prentice Hall, Englewood Cliffs, 3rd edition, 1991.

R. Isermann. *Digital Control Systems*, volume I. Fundamentals Deterministic Control. Springer-Verlag, 2nd edition, 1991a.

R. Isermann. *Digital Control Systems*, volume II. Stochastic Control, Multivariable Control, Adaptive Control, Applications. Springer-Verlag, 2nd edition, 1991b.

S.M. Kay. *Modern Spectral Estimation: Theory and Application.* Prentice Hall, 1988.

M. Kunt. *Traitement Numérique des Signaux.* Dunod, Paris, 1981.

H. Kwakernaak and R. Sivan. *Modern Signals and Systems.* Prentice Hall, Englewood Cliffs, 1991.

L. Ljung. *System Identification. Theory for the User.* Prentice Hall, Englewood Cliffs, 1987.

R.G. Lyons. *Understanding Digital Signal Processing.* Addison-Wesley, 1997.

J. Max. *Méthodes et Techniques du Traitement du Signal et Applications aux Mesures Physiques.* Masson, Paris, 1985.

R.H. Middleton and G.C. Goodwin. Improved finite word length characteristics in digital control using delta operator. *IEEE Trans. Automat. Control*, AC-31 (11):1015–1021, 1986.

R.H. Middleton and G.C. Goodwin. *Digital Control and Estimation.* Prentice Hall, Englewood Cliffs, 1990.

S.K. Mitra and J.F. Kaiser, editors. *Handbook for Digital Signal Processing.* Wiley, New York, 1993.

K. Ogata. *Discrete-Time Control Systems.* Prentice Hall, Englewood Cliffs, New Jersey, 1987.

A.V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, 1989.

S.J. Orfanidis. *Optimum Signal Processing.* Prentice Hall, 2nd edition, 1996.

J.G. Proakis and D.G. Manolakis. *Digital Signal Processing: Principles, Algorithms and Applications.* Macmillan, New York, 1996.

L.R. Rabiner and B. Gold. *Theory and Application of Digital Signal Processing.* Prentice Hall, 1975.

F. Roddier. *Distributions et Transformation de Fourier.* Ediscience, Paris, 1971.

C.B. Rorabaugh. *Digital Filter Designer's Handbook.* McGraw-Hill, 2nd edition, 1997.

W.P. Salman and M.S. Solotareff. *Le Filtrage Numérique.* Eyrolles, Paris, 1982.

D.E. Seborg, T.F. Edgar, and D.A. Mellichamp. *Process Dynamics and Control.* Wiley, New York, 1989.

Y. Sévely. *Systèmes et Asservissements Linéaires Echantillonnés.* Dunod, Paris, 1969.

T. Söderström and P. Stoica. *System Identification.* Prentice Hall, New York, 1989.

J.R. Treichler, C.R. Johnson, and M.G. Lawrence. *Theory and Design of Adaptive Filters.* Wiley, New York, 1987.

P.P. Vaidyanathan. *Multirate Systems and Filter Banks.* Prentice Hall, 1993.

Chapter 10

Identification Principles

The problems of identification of a linear system are discussed during three successive chapters: Chap. 10 presents, at a general level, non-parametric and parametric identification with the calculation of a predictor; Chap. 11 presents models and methods for parametric identification and Chap. 12 presents the algorithms of parametric identification.

10.1 System Description

The concerned systems are supposed to be linear and time-invariant.

10.1.1 System without Disturbance

The linear system of input $u(t)$, of output $y(t)$, which is time-invariant and causal, can be described by its impulse response $g(k)$ such that

$$y(t) = \sum_{k=1}^{\infty} g(k)\, u(t-k) \qquad , \quad t = 0, 1, 2, \ldots \tag{10.1}$$

The sampling instants are denoted by $0, 1, 2 \ldots$ as if they were separated by a unit sampling period T_s in order to simplify the notation.

Remark:

Note that $y(t)$ depends on $u(t-1)$, $u(t-2)$, ..., but not on $u(t)$, because it is estimated that the output is not immediately influenced by the input, even if the system presents no time delay.

By introducing the delay operator q^{-1} such that

$$q^{-1} y(t) = y(t-1) \tag{10.2}$$

the output is

$$\begin{aligned} y(t) &= \sum_{k=1}^{\infty} g(k)\, u(t-k) = \sum_{k=1}^{\infty} g(k)\, q^{-k}\, u(t) \\ &= \left[\sum_{k=1}^{\infty} q^{-k}\, g(k)\right] u(t) = G(q)\, u(t) \end{aligned} \tag{10.3}$$

The transfer function of the linear system is

$$G(q) = \sum_{k=1}^{\infty} q^{-k}\, g(k) \tag{10.4}$$

so that the output can be written in the absence of disturbance as

$$y(t) = G(q)\, u(t) \tag{10.5}$$

A linear filter $G(q)$ is strictly stable if

$$\sum_{k=1}^{\infty} k|g(k)| < \infty \tag{10.6}$$

and stable if

$$\sum_{k=1}^{\infty} |g(k)| < \infty \tag{10.7}$$

10.1.2 Disturbance Representation

A system is subjected to disturbances, so that the output can never be calculated even if the input is known. The disturbances may come from the measurement noise, from uncontrolled inputs. They are represented by simply adding a term to the output

$$y(t) = \sum_{k=1}^{\infty} g(k)\, u(t-k) + v(t) \tag{10.8}$$

according to the block diagram in Fig. 10.1.
The output signal $y(t)$ of the system is composed of a deterministic part related to the input and a stochastic part $v(t)$ related to the disturbances. In essence, the disturbance cannot be predicted and is often described by means of a sequence $\{e(t)\}$ of independent random variables having some given probability density. The disturbance model is then written as

$$v(t) = \sum_{k=0}^{\infty} h(k)\, e(t-k) \tag{10.9}$$

with the frequent hypothesis: $h(0) = 1$.

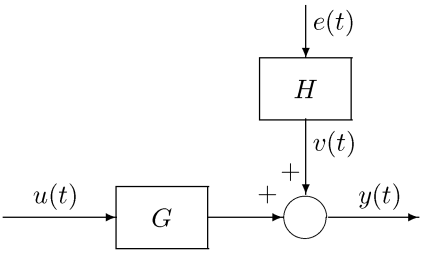


Figure 10.1: System with disturbance

Often, the sequence $\{e(t)\}$ is specified only by its mean and its variance. The mean is equal to the expectation of the variable

$$\mu_v = \mathrm{E}[v(t)] = \sum_{k=0}^{\infty} h(k)\ \ \mathrm{E}[e(t-k)] \tag{10.10}$$

while the covariance is equal to

$$\begin{aligned} \mathrm{E}[v(t)\, v(t-\tau)] &= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} h(k)\, h(j)\, \mathrm{E}[e(t-k)\, e(t-\tau-j)] \\ &= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} h(k)\, h(j) \delta(k-\tau-j) \lambda^2 \\ &= \lambda^2 \sum_{k=0}^{\infty} h(k)\, h(k-\tau) \end{aligned} \tag{10.11}$$

where λ^2 is the variance of e. Of course, it is assumed that

$$h(i) = 0 \quad \text{if } i < 0 \tag{10.12}$$

Due to the disturbance, $y(t)$ is a random variable such that

$$\mathrm{E}[y(t)] = G(q)\, u(t) + \mathrm{E}[v(t)] \tag{10.13}$$

In the presence of the disturbance, the output is

$$y(t) = G(q)\, u(t) + H(q)\, e(t) \tag{10.14}$$

$e(t)$ is often chosen as white noise (its spectral density is constant in all the frequency domain). The name of white noise comes from the analogy with white light, which contains all frequencies. $e(t)$ is a sequence of independent random variables of zero mean and variance λ^2.

10.2 Non-Parametric Identification

10.2.1 Frequency Identification

As well as in continuous time, where the Bode and Nyquist diagrams allow us to characterize the systems and discuss their stability, it is possible, in discrete

time, to subject the considered system without the disturbance to a sinusoidal input

$$u(t) = \cos(\omega t) = \mathcal{R}e\left[\exp(j\omega t)\right] \ \forall t \geq 0 \tag{10.15}$$

The output is then

$$\begin{aligned} y(t) &= \sum_{k=1}^{\infty} g(k)\mathcal{R}e\left[\exp(j\omega(t-k))\right] = \mathcal{R}e\left[\sum_{k=1}^{\infty} g(k)\exp(j\omega(t-k))\right] \\ &= \mathcal{R}e\left[\exp(j\omega t)\sum_{k=1}^{\infty} g(k)\exp(-j\omega k)\right] = \mathcal{R}e\left[\exp(j\omega t)G(\exp(j\omega))\right] \\ &= |G(\exp(j\omega))|\cos(\omega t + \phi) \end{aligned} \tag{10.16}$$

where ϕ is the phase shift of the output with respect to the input.

If the signal $u(t)$ is causal ($u(t) = 0 \ \forall t < 0$), the output is equal to

$$y(t) = \mathcal{R}e\left[\exp(j\omega t)\sum_{k=1}^{t} g(k)\exp(-j\omega k)\right] \tag{10.17}$$

Thus, it is theoretically possible to determine the discrete transfer function by this frequency approach (Ljung, 1987).

This elementary approach can be improved (Söderström and Stoica, 1989) by multiplying the output by $\cos(\omega t)$ on one hand giving a signal y_c, and by $\sin(\omega t)$ on the other hand giving a signal y_s. These signals are integrated over a period T, which allows us to decrease the noise influence, and results in

$$\begin{cases} \displaystyle\int_0^T y(t)\cos(\omega t)dt \approx \frac{bT}{2}\sin(\phi) \\ \displaystyle\int_0^T y(t)\sin(\omega t)dt \approx \frac{bT}{2}\cos(\phi) \end{cases} \tag{10.18}$$

by setting $y(t) = b\cos(\omega t + \phi)$.

10.2.2 Identification by Correlation Analysis

From Eq. (10.8), representing the system in the presence of a disturbance, it is possible to get the relation between the correlation functions

$$R_{yu}(n) = \mathrm{E}[y(t)u(t-\tau)] = \sum_{i=1}^{\infty} g(i)R_{uu}(n-i) \tag{10.19}$$

In fact, the correlation functions are calculated from the data according to the relations (9.72) and (9.73).

Practically, in relation (10.19), the correlation functions R_{yu}, R_{uu} and the coefficients $g(i)$ of the impulse response are replaced by their estimations. By knowing the correlation functions, this amounts to solving a system of infinite dimension with respect to the coefficients $g(i)$. Moreover, the sequence $g(i)$ is

truncated above a given threshold M, so that the linear system to be solved becomes

$$\hat{R}_{yu}(n) = \sum_{i=1}^{M} \hat{g}(i)\hat{R}_{uu}(n-i) \ , \quad n = 1, \ldots, M \tag{10.20}$$

Eventually, the system can be solved according to a least-squares method for a number of equations larger than M.

This method thus provides an estimation of the system impulse response corresponding to a finite impulse response (FIR) filter.

Example 10.1: **Identification by Correlation Analysis for a Chemical Reactor**

The chemical reactor described in Chap. 19 has been subjected to a pseudo-random binary sequence (Fig. 12.14). A sampling period $T_s = 25$ s was retained. The identification by correlation analysis was realized by means of Matlab©. The impulse response thus obtained is represented (Fig. 10.2) on the same graph as the impulse response from the ARMAX model obtained by parameter identification ($na = 2, nb = 2, nc = 2$). The order of magnitude is more or less respected, but the quality of the identification by correlation analysis is not very good. The 'o' points inside the confidence interval (dashed line) for the correlation analysis are neglected.

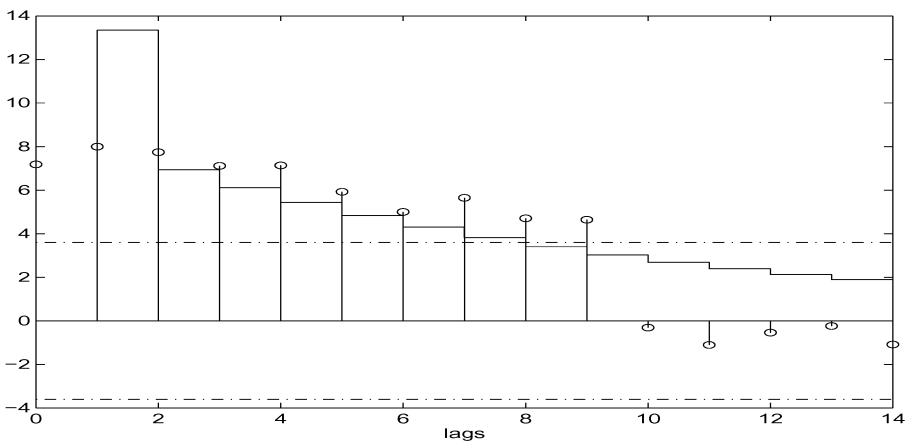


Figure 10.2: Identification by correlation analysis with the data coming from the chemical reactor simulated with a measurement noise: impulse response (symbol 'o') obtained with $M = 15$ compared to the impulse response obtained according to an ARMAX model by parameter identification (stairs curve)

10.2.3 Spectral Identification

A linear time-invariant system constitutes a linear filter. The signal crossing this filter is modified in particular with respect to frequency. Consider a strictly stable system $G(q)$ having as input $u(t)$ and output $y(t)$

$$y(t) = G(q)u(t) \tag{10.21}$$

The input $u(t)$ is unknown for $t < 0$ but assumed bounded for all t

$$|u(t)| < C_u \quad \forall t \tag{10.22}$$

The discrete Fourier transforms of the input and the output are

$$U_N(\omega) = \sum_{n=1}^{N} u(n) \exp(-j\omega n) \quad \text{and} \quad Y_N(\omega) = \sum_{n=1}^{N} y(n) \exp(-j\omega n) \tag{10.23}$$

According to Eq. (10.8), provided that the input and the disturbance are independent, the spectral and cross-spectral densities are related by the two following equations

$$\begin{cases} \Phi_{yy}(\omega) = |G(\exp(j\omega))|^2 \Phi_{uu}(\omega) + \Phi_{vv}(\omega) \\ \Phi_{yu}(\omega) = G(\exp(j\omega)) \Phi_{uu}(\omega) \end{cases} \tag{10.24}$$

The transfer function equal to

$$\hat{G}(\omega) = \frac{Y_N(\omega)}{U_N(\omega)} \tag{10.25}$$

thus can be estimated in the frequency domain as

$$\hat{G}(\omega) = \frac{\hat{\Phi}_{yu}(\omega)}{\hat{\Phi}_{uu}(\omega)} \tag{10.26}$$

with the cross-spectral and spectral densities respectively and rigorously equal to

$$\begin{cases} \Phi_{yu}(\omega) = \dfrac{1}{2\pi} \displaystyle\sum_{n=-\infty}^{\infty} R_{yu}(n) \exp(-j\omega n) \\ \Phi_{uu}(\omega) = \dfrac{1}{2\pi} \displaystyle\sum_{n=-\infty}^{\infty} R_{uu}(n) \exp(-j\omega n) \end{cases} \tag{10.27}$$

In fact, the cross-spectral and spectral densities are, respectively, estimated as

$$\begin{cases} \hat{\Phi}_{yu}(\omega) = \dfrac{1}{2\pi} \displaystyle\sum_{n=-N}^{N} \hat{R}_{yu}(n) \exp(-j\omega n) = \dfrac{1}{2\pi N} Y_N(\omega) U_N(-\omega) \\ \hat{\Phi}_{uu}(\omega) = \dfrac{1}{2\pi} \displaystyle\sum_{n=-N}^{N} \hat{R}_{uu}(n) \exp(-j\omega n) = \dfrac{1}{2\pi N} |U_N(\omega)|^2 \end{cases} \tag{10.28}$$

This approach gives poor results (Söderström and Stoica, 1989) because the estimation of the cross-correlation is bad when n becomes large. For this reason, it is better to apply a window so that these terms $\hat{R}_{yu}(n)$ and $\hat{R}_{uu}(n)$ (with n large) are filtered and become negligible. For example, the filtered cross-spectral density becomes

$$\hat{\Phi}_{yu}(\omega) = \frac{1}{2\pi} \sum_{n=-N}^{N} \hat{R}_{yu}(n) w(n) \exp(-j\omega n) \tag{10.29}$$

This equation can also be used in the case of autocorrelations. The window is characterized by the vector w; among different windows, cite the rectangle window (Fig. 10.3) such that

$$w(n) = \begin{cases} 1 & \text{if } |n| \leq M \\ 0 & \text{if } |n| > M \end{cases} \tag{10.30}$$

and the Hamming window (Fig. 10.3) such that

$$w(n) = \begin{cases} 0.54 - 0.46 \cos(2\pi \dfrac{n}{M-1}) & \text{if } |n| \leq M \\ 0 & \text{if } |n| > M \end{cases} \tag{10.31}$$

Many other types of windows exist (triangular, Bartlett, Blackman, Chebyshev, Hanning, Kaiser, etc.). With the infinite sequence being truncated by the use of the window, the Gibbs phenomenon (oscillations) may occur in a more or less important manner according to the type of window and the choice of this window width.

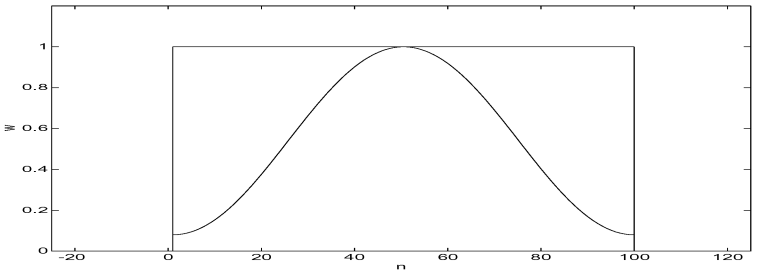


Figure 10.3: Hamming and rectangle windows

Example 10.2: **Influence of the Window on the Spectral Density**

The signal $x(t) = \cos(2\pi 20t) + 0.5e(t)$ has been sampled at a frequency of 200 Hz and considered over 1024 measurement points; $e(t)$ is a random number uniformly distributed in $[0, 1]$; thus the signal is noisy. The spectral density of this signal obviously makes appear the 20 Hz frequency of the signal (Fig. 10.4).

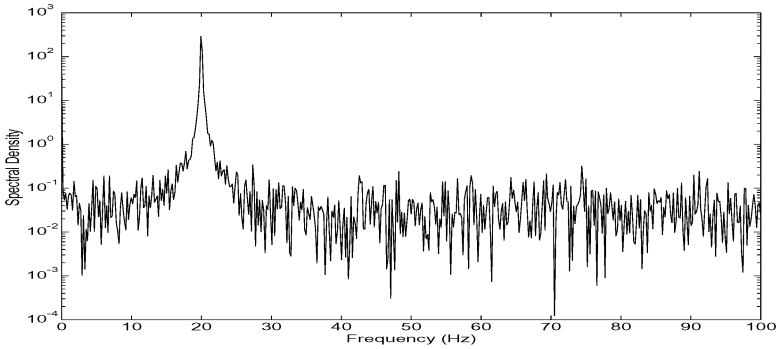


Figure 10.4: Spectral density of the signal without window

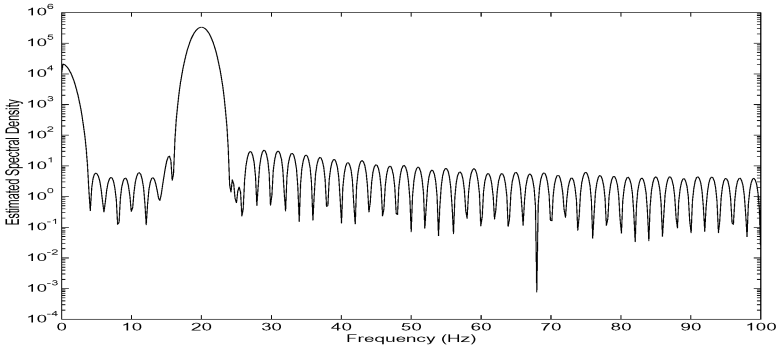


Figure 10.5: Spectral density of the signal with Hamming window of width $M = 100$

Equation (10.29) has been applied to the autocorrelation function of the signal by multiplying the signal by a Hamming window for which $M = 100$. The spectral density thus estimated is shown in Fig. 10.5. The fundamental frequency of 20 Hz appears clearly. The influence of the choice of M can lead to relatively different results.

By estimating the ratio according to Eq. (10.26) of the spectral densities in this manner, it is possible to estimate the transfer function of the considered system.

Example 10.3: **Spectral Identification for a Chemical Reactor**

The same data as in correlation analysis (Sect. 10.2.2) concerning the chemical reactor described in Chap. 19 are used. The sampling period is $T_s = 25$ s. The spectral identification has been performed by means of Matlab©. The Bode diagram of the transfer function obtained by spectral identification is rather close

(Fig. 10.6) to that of the transfer function obtained by parameter identification with an ARMAX model ($na = 2, nb = 2, nc = 2$), especially at low frequencies where the modulus is large. Note the maximum frequency: $\omega = \pi/T_s \approx 0,125$.

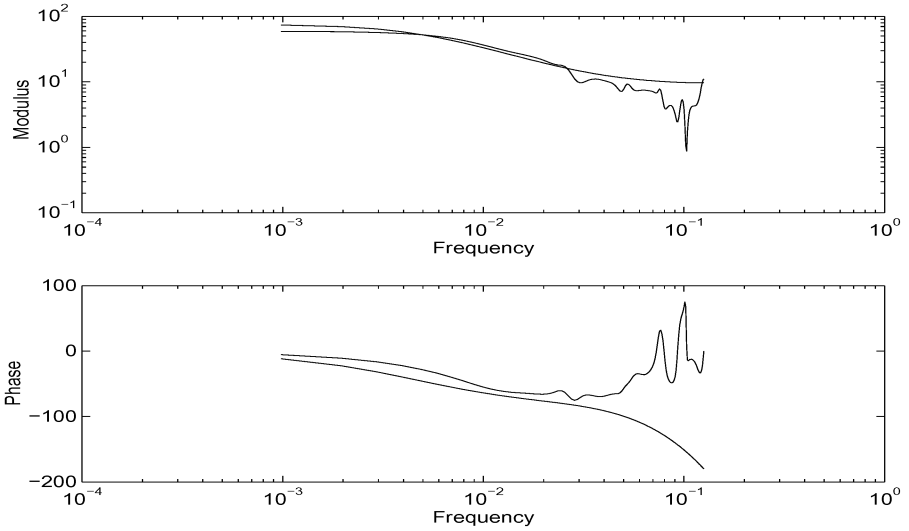


Figure 10.6: Spectral identification for the data coming from the chemical reactor simulated with measurement noise: comparison of Bode diagrams of the transfer functions obtained by spectral identification and by an ARMAX model ($na = 2, nb = 2, nc = 2$) (the diagram of the ARMAX model is smooth)

10.3 Parametric Identification

10.3.1 Prediction Principles

In this part, general principles of prediction (Ljung, 1987) are described. In the following chapters, they will be largely used in parameter identification. The system is assumed to be described by the following model

$$y(t) = G(q)u(t) + H(q)e(t) \tag{10.32}$$

with polynomials $G(q)$ and $H(q)$

$$G(q) = \sum_{i=1}^{\infty} g(i)q^{-i}\,, \quad H(q) = \sum_{i=0}^{\infty} h(i)q^{-i} \tag{10.33}$$

The q notation of the filters (e.g. $G(q)$) facilitates the discussions about analyticity of the corresponding function depending on z (e.g. $G(z)$). Qualifying

the two previous functions as a q polynomial is certainly improper, as these are polynomials with respect to q^{-1} and only functions with respect to z can be considered as analytic. However, we will often keep this practical designation.

Concerning simulation, when the previous system was subjected to an input u_s without disturbance, the resulting output would be equal to

$$y_s(t) = G(q)u_s(t) \tag{10.34}$$

Similarly, the influence of the disturbance could be simulated by creating a random sequence $e_s(t)$, which would be white noise such that the disturbance is

$$v_s(t) = H(q)e_s(t) \tag{10.35}$$

Then it is possible to know the system response to both input and disturbance.

10.3.2 One-Step Prediction

Disturbance Prediction

Given the model of the disturbance as

$$v(t) = H(q)e(t) = \sum_{i=0}^{\infty} h(i)e(t-i) \tag{10.36}$$

and the associated filter $H(q)$ which is assumed to be stable, as

$$\sum_{i=0}^{\infty} |h(i)| < \infty \tag{10.37}$$

a key-problem in identification is to be able to calculate the sequence $e(t)$ whose statistical properties are known (this is white noise). Set

$$e(t) = H_{inv}(q)v(t) = \sum_{i=0}^{\infty} h_{inv}(i)v(t-i) \tag{10.38}$$

with the analogous stability condition

$$\sum_{i=0}^{\infty} |h_{inv}(i)| < \infty \tag{10.39}$$

Let the function $H(z)$ be equal to

$$H(z) = \sum_{i=0}^{\infty} h(i)z^{-i} \tag{10.40}$$

and assume its inverse $1/H(z)$ to be analytical in the domain $|z| \geq 1$, the function of which is defined by

$$\frac{1}{H(z)} = H_{inv}(z) = \sum_{i=0}^{\infty} h_{inv}(i)z^{-i} \tag{10.41}$$

which means that the filter $H_{inv}(q)$ is stable. The coefficients $h_{inv}(i)$ are those searched for. Then the linear system or following filter is defined

$$H^{-1}(q) = \sum_{i=0}^{\infty} h_{inv}(i) q^{-i} \tag{10.42}$$

Ljung (1987) shows that $H_{inv}(q) = H^{-1}(q)$ satisfies Eq. (10.38). The coefficients of the filter H_{inv} can be calculated from Eq. (10.41). The fact that $H(q)$ is an inversely stable filter implies that the function $1/H(z)$ has no poles on or outside the unit circle, in the same way as $H(z)$ has no zeros on or outside the unit circle.
Remark: the calculation of the filter coefficients necessitates us to resort to the analytical functions with respect to z.

Example 10.4: **Inverse for a Moving Average Filter**
Ljung (1987) considers the following disturbance model

$$v(t) = e(t) + c_1 e(t-1) \tag{10.43}$$

corresponding to a first-order moving average model, which can be described by the following filter $H(q)$

$$H(q) = 1 + c_1 q^{-1} \tag{10.44}$$

The associated function is

$$H(z) = 1 + c_1 z^{-1} = \frac{z + c_1}{z} \tag{10.45}$$

This function, having a pole at $z = 0$ and a zero at $z = -c_1$, is analytical provided that $|c_1| < 1$ (no poles, nor zeros on or outside the unit circle). In this case, following the previous method, the inverse filter can be described by

$$H^{-1}(z) = \frac{1}{1 + c_1 z^{-1}} = \sum_{i=0}^{\infty} (-c_1 z^{-1})^i = \sum_{i=0}^{\infty} (-c_1)^i z^{-i} \tag{10.46}$$

and the sequence $e(t)$ results as

$$e(t) = \sum_{i=0}^{\infty} (-c_1)^i v(t-i) \tag{10.47}$$

In the case of one-step prediction, the signal $v(t)$ is assumed to be known up to time $t-1$, and we wish to predict the value of $v(t)$, thus, one step ahead. $v(t)$ is decomposed into two parts, one unknown, the other one known

$$v(t) = \sum_{i=0}^{\infty} h(i) e(t-i) = h(0) e(t) + \sum_{i=1}^{\infty} h(i) e(t-i) = e(t) + \sum_{i=1}^{\infty} h(i) e(t-i) \tag{10.48}$$

with $h(0) = 1$ (H is monic). The sum term of the right-member is known, as it uses only values of $e(t)$ included between 0 and $t-1$. Note that

$$w(t-1) = \sum_{i=1}^{\infty} h(i)e(t-i) \tag{10.49}$$

so that

$$v(t) = e(t) + w(t-1) \tag{10.50}$$

Denote by $\hat{v}(t|t-1)$ the prediction of $v(t)$ realized from the information from instant 0 until $(t-1)$. $\hat{v}(t|t-1)$ is the expectation of $v(t)$

$$\hat{v}(t|t-1) = E[v(t)] \tag{10.51}$$

hence

$$\hat{v}(t|t-1) = E[e(t)] + w(t-1) \tag{10.52}$$

It follows that the stochastic character of the prediction $v(t)$ depends only on the statistical properties of $e(t)$. Suppose that the random sequence $e(t)$ is uniformly distributed, thus it is white noise. As the mean of $e(t)$ is zero, the prediction of $v(t)$ is equal to

$$\hat{v}(t|t-1) = w(t-1) = \sum_{i=1}^{\infty} h(i)e(t-i) \tag{10.53}$$

This prediction minimizes the variance of the prediction error (Ljung, 1987), thus providing the optimal predictor $x(t)$

$$\min_{x(t)} \mathrm{E}(v(t) - x(t))^2 \Longleftrightarrow x(t) = \hat{v}(t|t-1) \tag{10.54}$$

In fact, the expression (10.53) of the predictor is not adapted to calculation, as only the signals $v(i)$ are known until $t-1$. Thus, the expression (10.53) is transformed in order to make the known signals appear according to a classical method in identification

$$\begin{aligned} \hat{v}(t|t-1) &= \left[\sum_{i=1}^{\infty} h(i)q^{-i}\right] e(t) \\ &= \left[H(q) - 1\right] e(t) \\ &= \left[H(q) - 1\right] H^{-1}(q)\, v(t) \\ &= \left[1 - H^{-1}(q)\right]\, v(t) \end{aligned} \tag{10.55}$$

The coefficients of the filter $H^{-1}(q)$ will be calculated in the same way as those of the function $H^{-1}(z) = 1/H(z)$. It is possible to write the following equivalent form

$$H(q)\,\hat{v}(t|t-1) = \left[H(q) - 1\right]\, v(t) = \sum_{k=1}^{\infty} h(k)v(t-k) \tag{10.56}$$

Example 10.5: **Predictor of the Disturbance**

Consider again the previous model of disturbance (moving average of order 1)

$$v(t) = e(t) + c_1 e(t-1) \tag{10.57}$$

hence

$$H(q) = 1 + c_1\, q^{-1} \tag{10.58}$$

The prediction formula

$$H(q)\, \hat{v}(t|t-1) = [H(q) - 1]\; v(t) \tag{10.59}$$

thus gives

$$\hat{v}(t|t-1) + c_1\, \hat{v}(t-1|t-2) = c_1\, v(t-1) \tag{10.60}$$

or still

$$\begin{aligned} \hat{v}(t|t-1) &= \frac{1}{1 + c_1\, q^{-1}}\, c_1\, v(t-1) = \left[\sum_{i=0}^{\infty} (-c_1)^i q^{-i}\right] c_1\, v(t-1) \\ &= -\sum_{i=0}^{\infty} (-c_1)^{i+1}\, v(t-i-1) = -\sum_{i=1}^{\infty} (-c_1)^i\, v(t-i) \end{aligned} \tag{10.61}$$

Output Prediction

The input u is known at instant $t-1$ and at the previous instants (recall that $y(t)$ depends on $u(t-1)$, $u(t-2)$, ...); the output y is known until instant $t-1$; the output $y(t)$ is sought. With the input-output model being

$$y(t) = G(q)\, u(t) + v(t) \tag{10.62}$$

the right-member is composed of a deterministic part $G(q)u(t)$ and a stochastic part $v(t)$. The prediction of the output results

$$\hat{y}(t|t-1) = G(q)\, u(t) + \hat{v}(t|t-1) \tag{10.63}$$

thus by using the prediction of the disturbance

$$\begin{aligned} \hat{y}(t|t-1) &= G(q)\, u(t) + \left[1 - H^{-1}(q)\right]\; v(t) \\ &= G(q)\, u(t) + \left[1 - H^{-1}(q)\right]\; [y(t) - G(q)\, u(t)] \end{aligned} \tag{10.64}$$

hence, finally,

$$\hat{y}(t|t-1) = H^{-1}(q)\, G(q)\, u(t) + \left[1 - H^{-1}(q)\right]\; y(t) \tag{10.65}$$

or in a form making use only of $H(q)$

$$H(q)\, \hat{y}(t|t-1) = G(q)\, u(t) + [H(q) - 1]\; y(t) \tag{10.66}$$

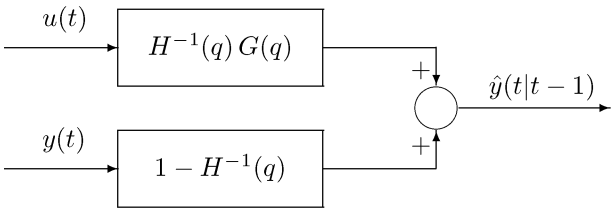


Figure 10.7: Filter of the predictor

The prediction of the output will be used in regulation; it acts as a linear filter having as inputs $u(t)$ and $y(t)$ and as an output the prediction $\hat{y}(t|t-1)$ (Fig. 10.7).
Remark:
Assuming that $G(z)$ has no poles on or outside the unit circle, and that $H(z)$ has no zeros on or outside the unit circle (stability of $1/H(z)$), and knowing that

$$G(z) = \sum_{i=1}^{\infty} g(i)\, z^{-i} \tag{10.67}$$

it is possible to write, by polynomial division, as an infinite (Laurent) expansion

$$\frac{G(z)}{H(z)} = \sum_{i=1}^{\infty} g_h(i)\, z^{-i} \tag{10.68}$$

On the other hand, recall that

$$H_{inv}(z) = H^{-1}(z) = \sum_{i=0}^{\infty} h_{inv}(i)\, z^{-i} \quad \text{with: } h_{inv}(0) = 1 \tag{10.69}$$

From the prediction Eq. (10.65), the predictor denoted by $\hat{f}(t)$ is rigorously written as

$$\begin{aligned} \hat{f}(t) &= H^{-1}(z)\, G(z)\, u(t) + \left[1 - H^{-1}(z)\right]\, y(t) \\ &= \sum_{i=1}^{\infty} g_h(i)\, u(t-i) - \sum_{i=1}^{\infty} h_{inv}(i)\, y(t-i) \end{aligned} \tag{10.70}$$

This expression assumes that the signals are known in the time interval $[-\infty, t-1]$, while, in fact, they are only known in $[0, t-1]$. The previous expression is then replaced by its approximation

$$\hat{y}(t|t-1) \approx \sum_{i=1}^{t} g_h(i)\, u(t-i) - \sum_{i=1}^{t} h_{inv}(i)\, y(t-i) \tag{10.71}$$

which amounts to replacing all the values corresponding to times larger than $t-1$ by 0. As the coefficients $g_h(i)$ and $h_{inv}(i)$ decrease quasi-exponentially, this approximation gives satisfactory results.

Prediction Error

The a priori prediction error is the difference between the output at time t and the prediction of this output made at time t from the previously known outputs, as

$$y(t) - \hat{y}(t|t-1) = -H^{-1}(q)\,G(q)\,u(t) + H^{-1}(q)\,y(t) \tag{10.72}$$

We notice that this prediction error corresponds to the non-predictible part of the output

$$y(t) - \hat{y}(t|t-1) = e(t) \tag{10.73}$$

The a priori prediction error is also called the innovation at time t.

10.3.3 p-Step Predictions

Disturbance Prediction

The problem is then to predict the disturbance $v(t+p)$, knowing $v(i)$ for $i \leq t$. The disturbance is related to white noise by the relation

$$v(t+p) = \sum_{i=0}^{\infty} h(i)\,e(t+p-i) \tag{10.74}$$

This sum can be decomposed into two parts, the second being known at time t, the first unknown but of zero mean

$$\begin{array}{rcl} v(t+p) & = & \displaystyle\sum_{i=0}^{p-1} h(i)\,e(t+p-i) + \sum_{i=p}^{\infty} h(i)\,e(t+p-i) \\ & = & H_i(q)e(t+p) + H_c(q)e(t) \end{array} \tag{10.75}$$

by defining the partial sums

$$H_i(q) = \sum_{i=0}^{p-1} h(i)\,q^{-i} \quad \text{and: } H_c(q) = \sum_{i=p}^{\infty} h(i)\,q^{p-i} \tag{10.76}$$

The expectation $v(t+p)$ is then equal to the known part

$$\hat{v}(t+p|t) = \sum_{i=p}^{\infty} h(i)\,e(t+p-i) = H_c(q)\,e(t) \tag{10.77}$$

thus giving the expression of the p-step prediction by using the relation $e(t) = H^{-1}(q)\,v(t)$

$$\hat{v}(t+p|t) = H_c(q)\,H^{-1}(q)\,v(t) \tag{10.78}$$

Output Prediction

The p-step output $y(t+p)$ is given by the relation

$$y(t+p) = G(q)u(t+p) + v(t+p) \tag{10.79}$$

With the inputs u being known from $-\infty$ to $t+p-1$ and the outputs y from $-\infty$ to t, we obtain

$$\begin{aligned}
\hat{y}(t+p|t) &= G(q)\,u(t+p) + \hat{v}(t+p|t) \\
&= G(q)\,u(t+p) + H_c(q)\,H^{-1}(q)\,v(t) \\
&= G(q)\,u(t+p) + H_c(q)\,H^{-1}(q)\;[y(t) - G(q)\,u(t)] \\
&= H_c(q)\,H^{-1}(q)\,y(t) + G(q)\;\left[1 - q^{-p}\,H_c(q)\,H^{-1}(q)\right]\,u(t+p)
\end{aligned} \tag{10.80}$$

Prediction Error

The a priori p-step prediction error for the output is equal to

$$\begin{aligned}
y(t+p) - \hat{y}(t+p|t) &= [1 - q^{-p}\,H_c(q)\,H^{-1}(q)][y(t+p) - G(q)\,u(t+p)] \\
&= [1 - q^{-p}\,H_c(q)\,H^{-1}(q)]\,v(t+p) \\
&= [1 - q^{-p}\,H_c(q)\,H^{-1}(q)]\,H(q)\,e(t+p) \\
&= [H(q) - q^{-p}\,H_c(q)]\,e(t+p) \\
&= H_i(q)\,e(t+p)
\end{aligned} \tag{10.81}$$

and thus depends on the successive noises $e(t+1), \ldots, e(t+p)$.

References

I.D. Landau. *Identification et Commande des Systèmes.* Hermès, Paris, 1988.

I.D. Landau. *System Identification and Control Design.* Prentice Hall, Englewood Cliffs, 1990.

I.D. Landau and A. Besançon Voda, editors. *Identification des Systèmes.* Hermès, Paris, 2001.

L. Ljung. *System Identification. Theory for the User.* Prentice Hall, Englewood Cliffs, 1987.

J. Richalet. *Pratique de l'Identification.* Hermès, Paris, 2nd edition, 1998.

T. Söderström and P. Stoica. *System Identification.* Prentice Hall, New York, 1989.

Chapter 11

Models and Methods for Parametric Identification

The estimation of the parameters of a given model is called parametric identification. For the notions of identifiability (the possibility of obtaining the parameters, for a certain number of experimental data) and of distinguishability (the possibility of distinguishing between distinct structures of models), the reader may refer to the following texts: Walter (1987); Walter and Pronzato (1997). The methods proposed in this chapter essentially concern time-invariant linear systems. Parametric identification is presented in the cases of transfer function and state-space systems.

11.1 Model Structure for Parametric Identification

It is necessary to choose a type of model before proceeding to the system identification. After the choice of the model structure, it will be possible to estimate the parameters of this model.

11.1.1 Linear Models of Transfer Functions

The output of the single-input single-output time-invariant linear system, subjected to a disturbance (Fig. 11.1), is modelled as

$$y(t) = G(q)\,u(t) + v(t) = G(q)\,u(t) + H(q)\,e(t) \tag{11.1}$$

This model is completely characterized by the coefficients of the impulse response $g(k)$, the spectral density of the disturbance $\Phi_v(\omega) = \lambda^2 |H(\exp(i\omega))|^2$, and the probability density function of the noise $e(t)$. The transfer functions are

$$G(q) = \sum_{i=1}^{\infty} g(i)\,q^{-i}\ , \quad H(q) = 1 + \sum_{i=1}^{\infty} h(i)\,q^{-i} \tag{11.2}$$

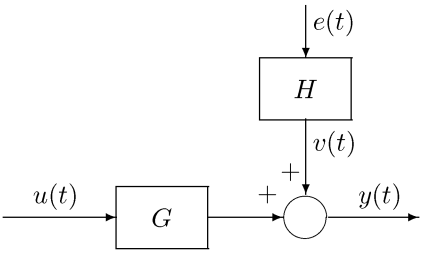


Figure 11.1: Single-input single-output linear system subjected to a disturbance

We notice that $H(q)$ is monic ($h(0) = 1$). In general, the number of coefficients in the expansion is taken to be finite. Moreover, the probability density function is not given, but, for example, it is Gaussian or its two first moments, the mean and the variance, are known

$$E[e(t)] = 0\ , \quad E[(e(t) - 0)^2] = \lambda^2 \tag{11.3}$$

The coefficients to be determined, in general, are not known through the knowledge of the physical model (although this can be very helpful in some cases), but by a black-box representation, thus demand estimation techniques and enter into the model as parameters to be determined. The parameter vector is denoted by θ and depends on the used identification method. The system model could be written in these conditions as

$$y(t) = G(q, \theta)\, u(t) + H(q, \theta)\, e(t) \tag{11.4}$$

and, in fact, covers a family of models dependent on θ. This general model is probabilistic, as the probability density function is specified through $e(t)$.

It is advisable to compare this model to the prediction model where the output depends only on the past inputs and outputs (cf. Eq. (10.65))

$$\hat{y}(t|t-1) = H^{-1}(q, \theta)\, G(q, \theta)\, u(t) + \left[1 - H^{-1}(q, \theta)\right]\, y(t) \tag{11.5}$$

Equation Error Models

Autoregressive exogenous ARX model.:
This model, which is the simplest, is written in the form of the following difference equation

$$y(t) + a_1\, y(t-1) + \ \ldots\ + a_{n_a}\, y(t - n_a) = b_1\, u(t-1) + \ \ldots\ + b_{n_b}\, u(t - n_b) + e(t) \tag{11.6}$$

or still

$$\begin{aligned} e(t) =& y(t) + a_1\, y(t-1) + \ \ldots\ + a_{n_a}\, y(t - n_a) - b_1\, u(t-1) - \ \ldots\ - b_{n_b}\, u(t - n_b) \\ =& A(q)\, y(t) - B(q)\, u(t) \end{aligned} \tag{11.7}$$

by setting the polynomials

$$A(q) = 1 + a_1\, q^{-1} + \ldots + a_{n_a}\, q^{-n_a} \quad \text{and: } B(q) = b_1\, q^{-1} + \ldots + b_{n_b}\, q^{-n_b} \tag{11.8}$$

In the expression of $B(q)$, we integrated the fact that the output is considered to be always delayed with respect to the input by at least a sampling period.

The error term $e(t)$ concerns the totality of the equation and appears as a moving average; for this reason, this type of model is also called an equation error model (Fig. 11.2).

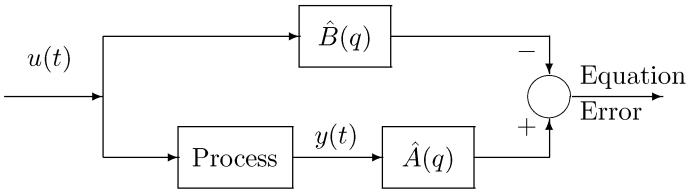


Figure 11.2: Principle of equation error

The parameter vector to be determined is

$$\theta = [a_1\,,\ \ldots\,,\ a_{n_a}\,,\ b_1\,,\ \ldots\,,\ b_{n_b}]^T \tag{11.9}$$

The model is thus of the form

$$A(q)\, y(t) = B(q)\, u(t) + e(t) \tag{11.10}$$

thus

$$y(t) = \frac{B(q)}{A(q)}\, u(t) + \frac{1}{A(q)}\, e(t) = G(q,\theta)\, u(t) + H(q,\theta)\, e(t) \tag{11.11}$$

giving the transfer functions G and H (Fig. 11.3)

$$G(q,\theta) = \frac{B(q)}{A(q)}\,, \quad H(q,\theta) = \frac{1}{A(q)} \tag{11.12}$$

The model and the disturbance thus possess the same dynamics, specified by the denominator $A(q)$.

The model is called ARX, as the part $A(q)\, y(t)$ is the regressive part in the expression of $y(t)$ and $B(q)\, u(t)$ is the exogenous part (external input).

The predictor associated with this model can be easily deduced from Eq. (11.5)

$$\hat{y}(t|\theta) = [1 - A(q)]\ y(t) + B(q)\, u(t) \tag{11.13}$$

The observation vector is

$$\phi(t) = [-y(t-1)\,,\ \ldots\,,\ -y(t-n_a)\,,\ u(t-1)\,,\ \ldots\,,\ u(t-n_b)]^T \tag{11.14}$$

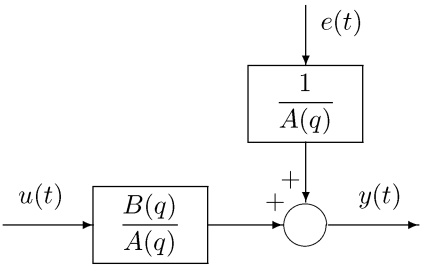


Figure 11.3: Structure of ARX equation error model: autoregressive eXogenous

The predictor of the output is thus the scalar product of the parameter vector by the observation vector

$$\hat{y}(t|\theta) = \phi^T(t)\,\theta = \theta^T\,\phi(t) \tag{11.15}$$

With the predictor being a linear function with respect to the parameters, the problem is a linear regression problem and the parameters can be searched by least-squares procedures.

ARMAX model: autoregressive moving average exogenous:

In order to better describe the disturbance term, rather than the ARX model, the ARMAX model is often preferred where the transfer function related to the disturbance will be the ratio of two polynomials according to the relation

$$y(t) = \frac{B(q)}{A(q)}\,u(t) + \frac{C(q)}{A(q)}\,e(t) = G(q,\theta)\,u(t) + H(q,\theta)\,e(t) \tag{11.16}$$

corresponding to the transfer functions G and H (Fig. 11.4)

$$G(q,\theta) = \frac{B(q)}{A(q)}\,, \quad H(q,\theta) = \frac{C(q)}{A(q)} \tag{11.17}$$

In the ARMAX model, as in the ARX model, the model and the disturbance possess the same dynamics, specified by the denominator $A(q)$.

The model expressed as a difference equation is thus

$$\begin{aligned} y(t) + a_1\,y(t-1) + \ldots + a_{n_a}\,y(t-n_a) = {} & b_1\,u(t-1) + \ldots + b_{n_b}\,u(t-n_b) \\ & + e(t) + c_1\,e(t-1) + \ldots + c_{n_c}\,e(t-n_c) \end{aligned} \tag{11.18}$$

with the polynomial

$$C(q) = 1 + c_1\,q^{-1} + \ldots + c_{n_c}\,q^{-n_c} \tag{11.19}$$

With respect to the ARX model, the parameter vector is increased by the coefficients $c_1, \ldots, c_{n_c}$ of $C(q)$.

To obtain the predictor of the output for this ARMAX model, it suffices to replace $e(t)$ in the previous model by $[y(t) - \hat{y}(t|\theta)]$, hence

$$\hat{y}(t|\theta) = \left[1 - \frac{A(q)}{C(q)}\right] y(t) + \frac{B(q)}{C(q)}\,u(t) \tag{11.20}$$

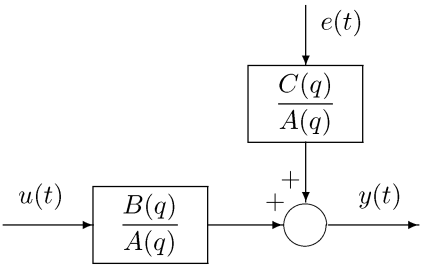


Figure 11.4: Structure of the equation error model ARMAX: autoregressive moving average exogenous

It is also possible to use the general predictor of Eq. (11.5). The previous expression, not being linear with respect to the sought parameters, is linearized by multiplying by $C(q)$ so as to obtain an equation of type (11.15)

$$C(q)\,\hat{y}(t|\theta) = [C(q) - A(q)]\;y(t) + B(q)\,u(t) \tag{11.21}$$

which amounts to filtering the input and the output by $C(q)$. In fact, the predictor is presented in the linear form deduced from the previous expression

$$\hat{y}(t|\theta) = \hat{y}(t|\theta) + [C(q) - A(q)]\;y(t) + B(q)\,u(t) - C(q)\,\hat{y}(t|\theta) \tag{11.22}$$

which is reordered according to the final linear form of the predictor

$$\begin{aligned} \hat{y}(t|\theta) &= [1 - A(q)]\;y(t) + B(q)\,u(t) + [C(q) - 1]\;[y(t) - \hat{y}(t|\theta)] \\ &= [1 - A(q)]\;y(t) + B(q)\,u(t) + [C(q) - 1]\;\epsilon(t,\theta) \end{aligned} \tag{11.23}$$

so as to make the prediction error intervene (Fig. 11.5). The error is called a priori if we take $\theta(t-1)$, a posteriori if we take $\theta(t)$

$$\epsilon(t,\theta) = y(t) - \hat{y}(t|\theta) \tag{11.24}$$

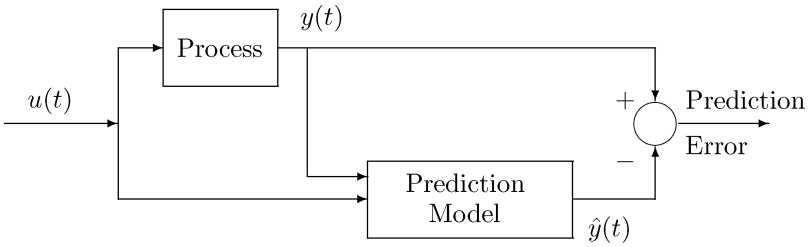


Figure 11.5: Principle of prediction error

We notice that the differences $C(q)-1$ and $1-A(q)$ make use of the coefficients c_i and a_i with $i \leq 1$. The observation vector is deduced from the prediction

model, which depends on the parameter vector through the prediction error $\epsilon(t, \theta)$

$$\begin{aligned} \phi(t,\theta) &= [-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b) , \\ &\quad \epsilon(t-1,\theta), \ldots, \epsilon(t-n_c,\theta)]^T \end{aligned} \tag{11.25}$$

and the parameter vector

$$\theta = [a_1 , \ldots , a_{n_a} , b_1 , \ldots , b_{n_b} , c_1 , \ldots , c_{n_c}]^T \tag{11.26}$$

which allows us to write the predictor in the classical form

$$\hat{y}(t|\theta) = \phi^T(t,\theta)\,\theta = \theta^T\,\phi(t,\theta) \tag{11.27}$$

This is then called a pseudo-linear regression (Ljung, 1987).

The ARMAX model itself includes an autoregressive term $A(q)y(t)$, an input term $B(q)u(t)$, a moving average term $C(q)e(t)$, and takes different forms according to the particular values of the coefficients:

- AR (autoregressive) model if $n_b = n_c = 0$. The output is expressed as a pure time series without any input signal

  $$A(q)\,y(t) = e(t) \tag{11.28}$$

  The associated predictor is equal to

  $$\hat{y}(t|\theta) = [1 - A(q)]\;y(t) \tag{11.29}$$

- MA (moving average) model if $n_a = n_b = 0$. The output is equal to

  $$y(t) = C(q)\,e(t) \tag{11.30}$$

  and does not depend on the input. The associated predictor is equal to

  $$\hat{y}(t|\theta) = [C(q) - 1]\;\epsilon(t|\theta) \tag{11.31}$$

- ARMA (autoregressive moving average) model if $n_b = 0$. The output is expressed as the relation

  $$A(q)\,y(t) = C(q)\,e(t) \tag{11.32}$$

  and simply describes the influence of a disturbance in a general manner. The associated predictor is equal to

  $$\hat{y}(t|\theta) = [C(q) - 1]\;\epsilon(t|\theta) + [1 - A(q)]\;y(t) \tag{11.33}$$

- ARIMA (autoregressive integrated moving average) model if $n_b = 0$ and if we force the factor $A(q)$ to contain as a factor an integrator term $(1 - q^{-1})$ (useful in suppressing offset in control). The output is expressed according to the relation

  $$A(q)\,y(t) = C(q)\,e(t) \tag{11.34}$$

  The influence of the disturbance, including this integrator, tends to describe a shift effect. The associated predictor is equal to

  $$\hat{y}(t|\theta) = [C(q) - 1]\;\epsilon(t|\theta) + [1 - A(q)]\;y(t) \tag{11.35}$$

- FIR (finite impulse response) model if $n_a = n_c = 0$. The output is simply equal to

$$y(t) = B(q)\, u(t) + e(t) \tag{11.36}$$

The associated predictor is equal to

$$\hat{y}(t|\theta) = B(q)\, u(t) \tag{11.37}$$

- ARX (AutoRegressive eXogenous) model if $n_c = 0$. The output is expressed according to the relation

$$A(q)\, y(t) = B(q)\, u(t) + e(t) \tag{11.38}$$

The associated predictor is equal to

$$\hat{y}(t|\theta) = [1 - A(q)]\ y(t) + B(q)\, u(t) \tag{11.39}$$

- ARARX model. This is obtained from an ARX model by replacing the error term taken as a moving average by an autoregressive error term. The ARARX model is thus written as

$$A(q)\, y(t) = B(q)\, u(t) + \frac{1}{D(q)}\, e(t) \tag{11.40}$$

The associated predictor is equal to

$$\hat{y}(t|\theta) = [1 - A(q)\, D(q)]\ y(t) + B(q) D(q)\, u(t) \tag{11.41}$$

- ARARMAX model. This is obtained by using an AutoRegressive Moving Average type (ARMA) for the equation error giving the equation

$$A(q)\, y(t) = B(q)\, u(t) + \frac{C(q)}{D(q)}\, e(t) \tag{11.42}$$

The ARARMAX model constitutes the more general case (Fig. 11.6) of the equation error models. The associated predictor is equal to

$$\hat{y}(t|\theta) = [C(q) - 1]\ \epsilon(t|\theta) + [1 - A(q)\, D(q)]\ y(t) + B(q)\, D(q)\, u(t) \tag{11.43}$$

All the models previously described include the transfer functions G and H, having the same polynomial $A(q)$ in their transfer functions. This may appear as a limitation. ARARX and ARARMAX models, moreover, include the polynomial $D(q)$ in the denominator of the transfer function relative to the error. For this reason, other types of models have been developed.

Output Error Models

Instead of having the error related to the equation as in the ARMAX model, it may seem more natural to introduce an error similar to a measurement error, related to the output y. The models, called output error models, will belong to this class (Fig. 11.7).

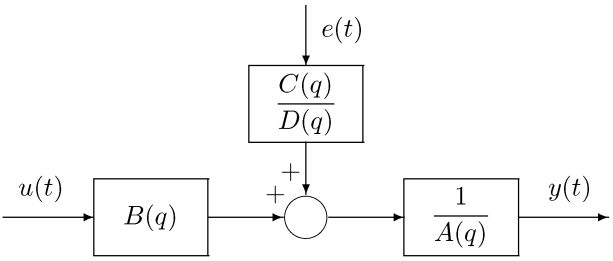


Figure 11.6: General equation error model

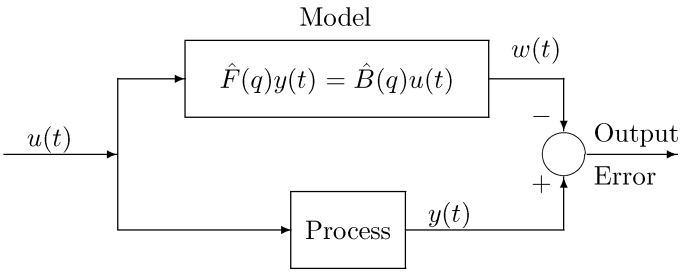


Figure 11.7: Principle of the output error

Elementary output error model:
The simplest output error model that can be developed will be in the form

$$y(t) = \frac{B(q)}{F(q)}\,u(t) + e(t) \tag{11.44}$$

where the error $e(t)$ bears only on the output $y(t)$, hence the name of this model (Fig. 11.8). In the ARX equation error model, the process transfer function was $B(q)/A(q)$; in this output error model, it is denoted by $B(q)/F(q)$ to distinguish the different roles played by the polynomials $A(q)$ and $F(q)$ in the model structure. If we call the output without error

$$w(t) = y(t) - e(t) \tag{11.45}$$

the difference equation associated with this model is written as

$$w(t)+f_1\,w(t-1)+\ldots+f_{n_f}\,w(t-n_f) = b_1\,u(t-1)+\ldots+b_{n_b}\,u(t-n_b) \tag{11.46}$$

Notice that $w(t)$ is an internal variable (Fig. 11.8), which is not observed and, in fact, depends on the parameter vector and thus will be denoted by $w(t,\theta)$. The predictor results immediately from the model

$$\hat{y}(t|\theta) = \frac{B(q)}{F(q)}\,u(t) = w(t,\theta) \tag{11.47}$$

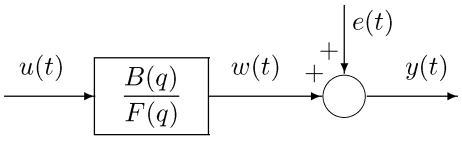


Figure 11.8: Elementary output error model

Taking as the parameter vector

$$\theta = [f_1\,,\,\ldots\,,\,f_{n_f}\,,\,b_1\,,\,\ldots\,,\,b_{n_b}]^T \tag{11.48}$$

and as the "observation" vector

$$\phi(t,\theta) = [-w(t-1,\theta),\,\ldots\,,\,-w(t-n_f,\theta),\,u(t-1),\,\ldots\,,\,u(t-n_b)]^T \tag{11.49}$$

the classical regression relation can be written

$$\hat{y}(t|\theta) = \phi^T(t,\theta)\,\theta = \theta^T\,\phi(t,\theta) \tag{11.50}$$

Box-Jenkins models:
The previous model can be improved by introducing a transfer function for white noise $e(t)$ so that we obtain

$$y(t) = \frac{B(q)}{F(q)}\,u(t) + \frac{C(q)}{D(q)}\,e(t) \tag{11.51}$$

By using the general equation of the predictor (11.5), we get the predictor associated with this model

$$\hat{y}(t|\theta) = \frac{D(q)\,B(q)}{C(q)\,F(q)}\,u(t) + \left[1 - \frac{D(q)}{C(q)}\right]\,y(t) \tag{11.52}$$

which could be transformed (refer to the general model) to lead to a pseudo-linear regression equation. By replacing $e(t)$ by $\hat{y}(t|\theta) - y(t)$, the same expression of the predictor would be obtained.

General Model for Identification

The most general model (Fig. 11.9) that can be written is

$$A(q)\,y(t) = \frac{B(q)}{F(q)}\,u(t) + \frac{C(q)}{D(q)}\,e(t) \tag{11.53}$$

This structure (Ljung, 1987) includes the different previous models, the equation error as well as the output error ones. Thus, if it covers all cases, it is not well adapted to each particular case. However, it is interesting to develop the identification methodology for the most general model. If, moreover, the process includes a time delay of n_r sampling periods between input u and output y, the model is modified by making the delay explicitly appear as

$$A(q)\,y(t) = q^{-n_r}\,\frac{B(q)}{F(q)}\,u(t) + \frac{C(q)}{D(q)}\,e(t) \tag{11.54}$$

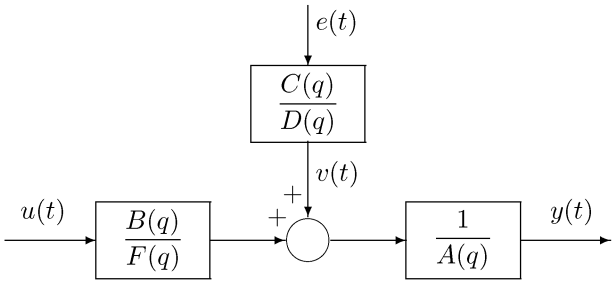


Figure 11.9: General model

the polynomials being the following

$$\begin{array}{lcl} A(q) & = & 1 + a_1\, q^{-1} + \ldots + a_{n_a}\, q^{-n_a} \\ B(q) & = & b_1\, q^{-1} + \ldots + b_{n_b}\, q^{-n_b} \\ C(q) & = & 1 + c_1\, q^{-1} + \ldots + c_{n_c}\, q^{-n_c} \\ D(q) & = & 1 + d_1\, q^{-1} + \ldots + d_{n_d}\, q^{-n_d} \\ F(q) & = & 1 + f_1\, q^{-1} + \ldots + f_{n_f}\, q^{-n_f} \end{array} \tag{11.55}$$

According to the general equation of the predictor (11.5), the expression of the predictor results

$$\hat{y}(t|\theta) = \left[\frac{D(q)\,B(q)}{C(q)\,F(q)}\right] u(t) + \left[1 - \frac{D(q)\,A(q)}{C(q)}\right] y(t) \tag{11.56}$$

This expression can be written in a recursive form by multiplying both members by $C(q)\,F(q)$. We then get a nonlinear equation with respect to the parameters

$$C(q)\,F(q)\,\hat{y}(t|\theta) = D(q)\,B(q)\,u(t) + F(q)\,\left[C(q) - D(q)\,A(q)\right]\,y(t) \tag{11.57}$$

From Eq. (11.56), the prediction error results

$$\epsilon(t,\theta) = y(t) - \hat{y}(t|\theta) = \frac{D(q)}{C(q)}\,\left[A(q)\,y(t) - \frac{B(q)}{F(q)}\,u(t)\right] \tag{11.58}$$

It is necessary to use auxiliary variables so that the final result is a pseudo-linear regression, that is, a linear equation with respect to the parameters that are to be determined. To lead to a simple form, it is necessary that the auxiliary variables allow the decoupling of the products of polynomials. Thus we successively introduce (each new variable depends on the known variables u, y or on the previous variables)

$$\begin{array}{lcl} w(t,\theta) & = & \dfrac{B(q)}{F(q)}\,u(t) \\ v(t,\theta) & = & A(q)\,y(t) - w(t,\theta) \end{array} \tag{11.59}$$

giving the prediction error

$$\epsilon(t,\theta) = \frac{D(q)}{C(q)}\,v(t,\theta) \tag{11.60}$$

The set of Eqs. (11.58), (11.60), (11.64), allows us to write the predictor

$$\left\{\begin{array}{ll}\hat{y}(t|\theta) & = y(t) - \epsilon(t,\theta) \\ & = [A(q)y(t) - A'^T \cdot Y] - [C(q)\epsilon(t,\theta) - C'^T \cdot E] \\ & = v(t,\theta) + w(t,\theta) - A'^T \cdot Y - D(q)v(t,\theta) + C'^T \cdot E \\ & = [v(t,\theta) - D(q)v(t,\theta)] + [F(q)w(t,\theta) - F'^T \cdot W] - A'^T \cdot Y + C'^T \cdot E \\ & = -D'^T \cdot V + B(q)u(t) - F'^T \cdot W - A'^T \cdot Y + C'^T \cdot E \\ & = -D'^T \cdot V + B'^T \cdot U - F'^T \cdot W - A'^T \cdot Y + C'^T \cdot E \end{array}\right. \tag{11.61}$$

leading to the usual linear form of regression

$$\hat{y}(t|\theta) = \theta^T \, \phi(t,\theta) \tag{11.62}$$

To obtain expression (11.61), the vectors ϕ and θ were symbolized by making use of internal sub-vectors

$$\phi = [-Y \, , \, U \, , \, -W \, , \, E \, , \, -V]^T \quad , \quad \theta = [A' \, , \, B' \, , \, F' \, , \, C' \, , \, D']^T \tag{11.63}$$

where Y, U, E, V, W are the vectors truncated from the values $y(t)$, $\epsilon(t)$, $v(t)$, $w(t)$, and A', F', C', D' are the vectors symbolizing the monic polynomials whose coefficient 1 of order 0 was taken off. B' is the vector symbolizing the polynomial B. Therefore, we can write the relations

$$\left\{\begin{array}{ll} A(q)y(t) & = A'^T \cdot Y + y(t) \\ B(q)u(t) & = B'^T \cdot U \\ C(q)\epsilon(t,\theta) & = C'^T \cdot E + \epsilon(t,\theta) \\ D(q)v(t,\theta) & = D'^T \cdot V + v(t,\theta) \\ F(q)w(t,\theta) & = F'^T \cdot W + w(t,\theta) \end{array}\right. \tag{11.64}$$

where $\cdot$ symbolizes the scalar product of the vectors.

The observation vector is then

$$\begin{array}{rcl}\phi(t,\theta) & = & [-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b), \\ & & -w(t-1,\theta), \ldots, -w(t-n_f,\theta), \epsilon(t-1,\theta), \ldots, \epsilon(t-n_c,\theta), \\ & & -v(t-1,\theta), \ldots, -v(t-n_d,\theta)]^T \end{array} \tag{11.65}$$

and the associated parameter vector is equal to

$$\theta = \left[a_1 \, , \, \ldots \, , \, a_{n_a} \, , \, b_1 \, , \, \ldots \, , \, b_{n_b} \, , \, f_1 \, , \, \ldots \, , \, f_{n_f} \, , \, c_1 \, , \, \ldots \, , \, c_{n_c} \, , \, d_1 \, , \, \ldots \, , \, d_{n_d}\right]^T \tag{11.66}$$

11.1.2 Models for Estimation in State Space

The advantage of the state-space representation coming from the fundamental equations is that it is closer to the physical process than the transfer function models. Thus, the parameter vector can be related to the physical parameters that govern the system. This may allow us to better control the variation domains of the parameters.

Discrete Kalman Filter

The process is represented in state space and discrete time for a multi-input multi-output system, taking into account the measurement noise $\boldsymbol{v}_k$ and the process noise $\boldsymbol{w}_k$ (concerning the states), by the following model

$$\begin{cases} \boldsymbol{x}_{k+1} & = \boldsymbol{A}_k\,\boldsymbol{x}_k + \boldsymbol{B}_k\,\boldsymbol{u}_k + \boldsymbol{G}_k\,\boldsymbol{w}_k \\ \boldsymbol{y}_k & = \boldsymbol{C}_k\,\boldsymbol{x}_k + \boldsymbol{v}_k \end{cases} \tag{11.67}$$

This model, in general, comes from the discretization of the continuous-time state-space model. The parameters of the model are the coefficients of matrices $\boldsymbol{A}_k$, $\boldsymbol{B}_k$, $\boldsymbol{C}_k$. The considered instants, denoted by subscript k, thus correspond to a sampling of period T_s. The noises $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ are supposed to be sequences of independent random variables of zero mean and of covariances

$$\begin{cases} \mathrm{E}\left[\boldsymbol{w}_k\,\boldsymbol{w}_l^T\right] & = \boldsymbol{Q}_k\delta_{kl} \\ \mathrm{E}\left[\boldsymbol{v}_k\,\boldsymbol{v}_l^T\right] & = \boldsymbol{R}_k\delta_{kl} \\ \mathrm{E}\left[\boldsymbol{w}_k\,\boldsymbol{v}_l^T\right] & = \boldsymbol{S}_k\delta_{kl} \end{cases} \tag{11.68}$$

where δ_{kl} is the Kronecker symbol ($= 0$ if $k \neq l$, $= 1$ if $k = l$). The matrices $\boldsymbol{Q}_k$ and $\boldsymbol{R}_k$ are symmetrical positive definite. Frequently, the measurement and process noises are assumed to be uncorrelated, so that in this case $\boldsymbol{S}_k = 0$. The case where these noises are correlated will be explained later.

The initial state $\boldsymbol{x}_0$ of the system must be specified, and is such that

$$\mathrm{E}\left[(\tilde{\boldsymbol{x}}_0)\,(\tilde{\boldsymbol{x}}_0)^T\right] = \boldsymbol{P}_0 \tag{11.69}$$

with $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \mathrm{E}(\boldsymbol{x})$. The matrix $\boldsymbol{P}_0$ is positive definite.

The noises $\boldsymbol{v}(t)$ and $\boldsymbol{w}(t)$ are any when the filter minimizes the a priori variance of the estimation error; they are assumed to be Gaussian when the filter maximizes the a posteriori probability of the variables to be estimated (Borne et al., 1990; Radix, 1970).

According to Eqs. (11.67,) which define a recurrence relation, it is possible to calculate the state and the output at instant k with respect to the initial conditions and the sequence of the inputs and the noises

$$\begin{cases} \boldsymbol{x}_k = & \Phi_{k,0}\boldsymbol{x}_0 + \displaystyle\sum_{i=0}^{k-1} \Phi_{k,i+1}[\boldsymbol{B}_i\boldsymbol{u}_i + \boldsymbol{G}_i w_i] \\ \boldsymbol{y}_k = & \boldsymbol{C}_k\,\boldsymbol{x}_k + \boldsymbol{v}_k \end{cases} \tag{11.70}$$

where $\Phi_{k,i}$ is the state transition matrix defined by

$$\Phi_{k,i} = \boldsymbol{A}_{k-1}\boldsymbol{A}_{k-2}\ldots\boldsymbol{A}_i \text{ if: } k \neq i\ ; \quad \Phi_{k,k} = \boldsymbol{I} \tag{11.71}$$

The Kalman filter (Kalman, 1960; Kalman and Bucy, 1961), allows us to calculate the prediction of $y(t)$ according to the equations

$$\begin{cases} \hat{\boldsymbol{x}}_{k+1} = & \boldsymbol{A}_k\,\hat{\boldsymbol{x}}_k + \boldsymbol{B}_k\,\boldsymbol{u}_k + \boldsymbol{K}_k\,[\boldsymbol{y}_k - \boldsymbol{C}_k\,\hat{\boldsymbol{x}}_k] \\ \hat{\boldsymbol{y}}_k = & \boldsymbol{C}_k\,\hat{\boldsymbol{x}}_k \end{cases} \tag{11.72}$$

where $\boldsymbol{K}_k$ is the Kalman gain matrix and $\boldsymbol{y}_k$ is the real measurement performed at instant k. This measurement is often denoted by $\boldsymbol{z}_k$ to distinguish the set $\{y\}$ of all the outputs from the set $\{z\}$ of the measured outputs.

In fact, the use of the Kalman filter is done iteratively in two stages

- Prediction stage:

$$\begin{array}{ll} \hat{\boldsymbol{x}}_{k+1|k} & = \boldsymbol{A}_k\,\hat{\boldsymbol{x}}_{k|k} + \boldsymbol{B}_k\,\boldsymbol{u}_k \\ \boldsymbol{P}_{k+1|k} & = \boldsymbol{A}_k\,\boldsymbol{P}_{k|k}\,\boldsymbol{A}_k^T + \boldsymbol{G}_k\boldsymbol{Q}_k\boldsymbol{G}_k^T \end{array} \tag{11.73}$$

- Correction stage (improvement of the estimation)

$$\begin{array}{ll} \hat{\boldsymbol{x}}_{k+1|k+1} & = \hat{\boldsymbol{x}}_{k+1|k} + \boldsymbol{K}_{k+1}\,(\boldsymbol{y}_{k+1} - \boldsymbol{C}_{k+1}\hat{\boldsymbol{x}}_{k+1|k}) \\ \boldsymbol{P}_{k+1|k+1} & = (\boldsymbol{I} - \boldsymbol{K}_{k+1}\boldsymbol{C}_{k+1})\boldsymbol{P}_{k+1|k}(\boldsymbol{I} - \boldsymbol{K}_{k+1}\boldsymbol{C}_{k+1})^T + \boldsymbol{K}_{k+1}\boldsymbol{R}_{k+1}\boldsymbol{K}_{k+1}^T \end{array} \tag{11.74}$$

Demonstrations:
1) By definition, the covariance matrix of the one-step predicted error is

$$\boldsymbol{P}_{k+1|k} \equiv \mathrm{E}[(\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}_{k+1|k})(\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}_{k+1|k})^T] \tag{11.75}$$

In the prediction stage, it results that

$$\begin{array}{ll} \boldsymbol{P}_{k+1|k} & = \mathrm{E}[\left\{\boldsymbol{A}_k(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k}) + \boldsymbol{G}_k\boldsymbol{w}_k\right\}\left\{\boldsymbol{A}_k(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k}) + \boldsymbol{G}_k\boldsymbol{w}_k\right\}^T] \\ & = \boldsymbol{A}_k\mathrm{E}[(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k})(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k})^T]\boldsymbol{A}_k^T + \boldsymbol{G}_k\mathrm{E}[\boldsymbol{w}_k\boldsymbol{w}_k^T]\boldsymbol{G}_k^T \\ & = \boldsymbol{A}_k\boldsymbol{P}_{k|k}\boldsymbol{A}_k^T + \boldsymbol{G}_k\boldsymbol{Q}_k\boldsymbol{G}_k^T \end{array} \tag{11.76}$$

which is indeed formula (11.73).
2) By definition, the covariance matrix of the filtered error is

$$\boldsymbol{P}_{k|k} \equiv \mathrm{E}[(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k})(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k})^T] \tag{11.77}$$

As

$$\begin{array}{ll} \hat{\boldsymbol{x}}_{k|k} & = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\,(\boldsymbol{y}_k - \boldsymbol{C}_k\hat{\boldsymbol{x}}_{k|k-1}) \\ & = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\,(\boldsymbol{C}_k\boldsymbol{x}_k + \boldsymbol{v}_k - \boldsymbol{C}_k\hat{\boldsymbol{x}}_{k|k-1}) \end{array} \tag{11.78}$$

it results that

$$\hat{\boldsymbol{x}}_{k|k} - \boldsymbol{x}_k = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{C}_k)(\hat{\boldsymbol{x}}_{k|k-1} - \boldsymbol{x}_k) + \boldsymbol{K}_k\boldsymbol{v}_k \tag{11.79}$$

giving the covariance matrix of the filtered error

$$\boldsymbol{P}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{C}_k)\boldsymbol{P}_{k|k-1}(\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{C}_k)^T + \boldsymbol{K}_k\boldsymbol{R}_k\boldsymbol{K}_k^T \tag{11.80}$$

which is indeed formula (11.74).

The gain matrix of the filter is calculated in order to minimize the following criterion

$$\mathrm{trace}(\boldsymbol{P}_{k|k}) \tag{11.81}$$

The Kalman filter is thus an optimal linear filter and the optimal gain minimizes the sum of the variances of the estimation errors, by definition of the trace[1]. The trace represents the confidence that we have in the previous estimation; the greater the confidence, the smaller the trace and the lower the gain. On the contrary, if we wish to take into account new measurements, the gain must not be too low. The gain thus represents, in fact, a compromise.

The minimization of the criterion gives the optimal gain matrix

$$\frac{\partial \text{trace}(\boldsymbol{P}_{k|k})}{\partial \boldsymbol{K}} = 0 \Longrightarrow \boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{C}_k^T \Sigma_k^{-1} \tag{11.82}$$

with

$$\Sigma_k = \boldsymbol{C}_k \boldsymbol{P}_{k|k-1} \boldsymbol{C}_k^T + \boldsymbol{R}_k \tag{11.83}$$

By use of the matrix inversion lemma [2], the following expression, which is frequently useful, is obtained

$$\Sigma_k^{-1} = \boldsymbol{R}_k^{-1} - \boldsymbol{R}_k^{-1}\,\boldsymbol{C}_k\,\boldsymbol{P}_{k|k}^{-1}\,\boldsymbol{C}_k^T \boldsymbol{R}_k^{-1} \tag{11.85}$$

By use of this expression of the optimal gain, the covariance matrix of the filtered error, cf. Eq. (11.74), can be simplified as

$$\boldsymbol{P}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{C}_k)\boldsymbol{P}_{k|k-1} \tag{11.86}$$

This relation has an important consequence

$$\text{trace}(\boldsymbol{P}_{k|k}) \leq \text{trace}(\boldsymbol{P}_{k|k-1}) \tag{11.87}$$

thus the trace decreases along the iterations.

The optimal gain can also be written again in the form

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k}\boldsymbol{C}_k^T \boldsymbol{R}_k^{-1} \tag{11.88}$$

thus as the covariance of the filtered error $(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k})$ becomes larger, the gain increases for a given error matrix $\boldsymbol{R}_k$.

Eq. (11.80) of the filtered error covariance matrix is called the Joseph form (Borne et al., 1990), and is more robust numerically (in particular, it guarantees that the matrix remains symmetrical positive definite along the iterations) than the simplified relation (11.86).

It is also possible to write the matrix inversion lemma

$$\boldsymbol{P}_{k|k}^{-1} = \boldsymbol{P}_{k|k-1}^{-1} + \boldsymbol{C}_k^T \boldsymbol{R}_k^{-1} \boldsymbol{C}_k \tag{11.89}$$

The Kalman filter can be considered in two different ways (Borne et al., 1990; Watanabe, 1992): as a filter or estimator if we first make the prediction

[1] The trace of a matrix is equal to the sum of the diagonal elements of the matrix. It is also equal to the sum of the eigenvalues of this matrix.

[2] The matrix inversion lemma concerning matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$, gives the following equality

$$(\boldsymbol{A} + \boldsymbol{B}\,\boldsymbol{C}\,\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\,\boldsymbol{B}\,(\boldsymbol{C}^{-1} + \boldsymbol{D}\,\boldsymbol{A}^{-1}\,\boldsymbol{B})^{-1}\,\boldsymbol{D}\,\boldsymbol{A}^{-1} \tag{11.84}$$

then the correction, or as a one-step ahead predictor if we first apply the correction then the prediction. The filter is currently used as an observer in the control laws, needing the knowledge of the states. The predictor is useful for studying the stability of the Kalman filter (Watanabe, 1992). In both cases, the prediction and correction formulae are identical. On the other hand, the explanation of the stages leads to two different sets of equations:
Kalman filter as an estimator:
Knowing the last estimation $\hat{\boldsymbol{x}}_{k|k}$, we first apply the prediction giving $\hat{\boldsymbol{x}}_{k+1|k}$, then the correction giving $\hat{\boldsymbol{x}}_{k+1|k+1}$, from which we deduce, after simplifications

$$\begin{aligned}
\hat{\boldsymbol{x}}_{k+1|k+1} &= (\boldsymbol{I} - \boldsymbol{K}_{k+1}\,\boldsymbol{C}_{k+1})\,(\boldsymbol{A}_k\,\hat{\boldsymbol{x}}_{k|k} + \boldsymbol{B}_k\,\boldsymbol{u}_k) + \boldsymbol{K}_{k+1}\,\boldsymbol{y}_{k+1}\ ;\ \hat{\boldsymbol{x}}_{0|0} = \hat{\boldsymbol{x}}_0 \\
\boldsymbol{\Sigma}_{k+1} &= \boldsymbol{R}_{k+1} + \boldsymbol{C}_{k+1}[\boldsymbol{A}_k\boldsymbol{P}_{k|k}\,\boldsymbol{A}_k^T + \boldsymbol{G}_k\boldsymbol{Q}_k\,\boldsymbol{G}_k^T]\boldsymbol{C}_{k+1}^T \\
\boldsymbol{K}_{k+1} &= [\boldsymbol{A}_k\,\boldsymbol{P}_{k|k}\,\boldsymbol{A}_k^T + \boldsymbol{G}_k\,\boldsymbol{Q}_k\,\boldsymbol{G}_k^T\,]\,\boldsymbol{C}_{k+1}^T\,\boldsymbol{\Sigma}_{k+1}^{-1} \\
\boldsymbol{P}_{k+1|k+1} &= [\boldsymbol{I} - \boldsymbol{K}_{k+1}\,\boldsymbol{C}_{k+1}]\,[\boldsymbol{A}_k\boldsymbol{P}_{k|k}\,\boldsymbol{A}_k^T + \boldsymbol{G}_k\boldsymbol{Q}_k\,\boldsymbol{G}_k^T]\ ; \quad \boldsymbol{P}_{0|0} = \boldsymbol{P}_0
\end{aligned} \tag{11.90}$$

Kalman filter as a one-step ahead predictor:
We first apply the prediction giving $\hat{\boldsymbol{x}}_{k+1|k}$ with respect to the estimation $\hat{\boldsymbol{x}}_{k|k}$, then we replace this estimation by the correction formula with respect to the previous prediction $\hat{\boldsymbol{x}}_{k|k-1}$, from which we obtain, after simplifications

$$\begin{aligned}
\hat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{A}_k\,\hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{B}_k\,\boldsymbol{u}_k + \boldsymbol{A}_k\,\boldsymbol{K}_k\,[\boldsymbol{y}_k - \boldsymbol{C}_k\,\hat{\boldsymbol{x}}_{k|k-1}]\ ; \qquad \hat{\boldsymbol{x}}_{0|-1} = \hat{\boldsymbol{x}}_0 \\
\boldsymbol{K}_k &= \boldsymbol{P}_{k|k-1}\,\boldsymbol{C}_k^T\,[\boldsymbol{C}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{C}_k^T + \boldsymbol{R}_k]^{-1}
\end{aligned} \tag{11.91}$$

with the covariance matrix of the predicted error solution of the Riccati discrete equation

$$\begin{aligned}
\boldsymbol{P}_{k+1|k} = {} & \boldsymbol{A}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{A}_k^T + \boldsymbol{G}_k\,\boldsymbol{Q}_k\,\boldsymbol{G}_k^T - \\
& \boldsymbol{A}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{C}_k^T\,[\boldsymbol{C}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{C}_k^T + \boldsymbol{R}_k]^{-1}\boldsymbol{C}_k\,\boldsymbol{P}_{k|k-1}^T\,\boldsymbol{A}_k^T\ ;\ \boldsymbol{P}_{0|-1} = \boldsymbol{P}_0
\end{aligned} \tag{11.92}$$

We notice in Eq. (11.91) that the gain of the predictor filter is equal to $\boldsymbol{A}_k\boldsymbol{K}_k$ which is thus the product of the state matrix $\boldsymbol{A}_k$ by the gain $\boldsymbol{K}_k$ of the estimator filter.

The case where the measurement and process noises are correlated, i.e. $\boldsymbol{S}_k \neq 0$, can be reduced to the simplified case $\boldsymbol{S}_k = 0$ by introducing the new model equivalent to the system (Watanabe, 1992)

$$\left\{\begin{aligned}
\boldsymbol{x}_{k+1} &= \boldsymbol{A}_k\,\boldsymbol{x}_k + \boldsymbol{B}_k\,\boldsymbol{u}_k + \boldsymbol{G}_k\,\boldsymbol{w}_k + \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}[\boldsymbol{y}_k - \boldsymbol{C}_k\,\boldsymbol{x}_k - \boldsymbol{v}_k] \\
&= \boldsymbol{A}'_k\,\boldsymbol{x}_k + \boldsymbol{B}_k\,\boldsymbol{u}_k + \boldsymbol{G}'_k\,\boldsymbol{y}_k + \boldsymbol{G}_k\,\boldsymbol{w}'_k \\
\boldsymbol{y}_k &= \boldsymbol{C}_k\,\boldsymbol{x}_k + \boldsymbol{v}_k
\end{aligned}\right. \tag{11.93}$$

with the new process noise $\boldsymbol{w}'_k$, which is not correlated with the measurement noise $\boldsymbol{v}_k$

$$\boldsymbol{w}'_k = \boldsymbol{G}_k\,\boldsymbol{w}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\boldsymbol{v}_k \tag{11.94}$$

and

$$\begin{aligned} \boldsymbol{A}'_k &= \boldsymbol{A}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\boldsymbol{C}_k \\ \boldsymbol{G}'_k &= \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1} \\ \mathrm{E}\left\{ \begin{bmatrix} \boldsymbol{w}'_k \\ \boldsymbol{v}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_l'^T & \boldsymbol{v}_l^T \end{bmatrix} \right\} &= \begin{bmatrix} \boldsymbol{Q}'_k & 0 \\ 0 & \boldsymbol{R}_k \end{bmatrix} \boldsymbol{\delta}_{kl} \\ \boldsymbol{Q}'_k &= \boldsymbol{Q}_k - \boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\boldsymbol{S}_k^T \end{aligned} \tag{11.95}$$

In these conditions, the Kalman filter can be solved as previously, but with this new model. From the equations of prediction (11.73) and correction (11.74), it is possible to notice that only the prediction stage is modified by the introduction of the new model.

Kalman filter as an estimator:

The formulae of the estimator result very simply from the formulae (11.90) of the estimator without correlation of the noises, by taking into account the new matrices $\boldsymbol{A}'_k$, $\boldsymbol{Q}'_k$ and the term $\boldsymbol{G}'_k y_k$.

$$\begin{aligned} \hat{\boldsymbol{x}}_{k+1|k+1} = \quad & (\boldsymbol{I} - \boldsymbol{K}_{k+1}\,\boldsymbol{C}_{k+1})\,(\boldsymbol{A}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{C}_k)\,\hat{\boldsymbol{x}}_{k|k} + \\ & (\boldsymbol{I} - \boldsymbol{K}_{k+1}\,\boldsymbol{C}_{k+1})\,\boldsymbol{B}_k\,\boldsymbol{u}_k + \\ & (\boldsymbol{I} - \boldsymbol{K}_{k+1}\,\boldsymbol{C}_{k+1})\boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{y}_k + \boldsymbol{K}_{k+1}\,\boldsymbol{y}_{k+1}\;; \qquad \hat{\boldsymbol{x}}_{0|0} = \hat{\boldsymbol{x}}_0 \\ \boldsymbol{\Sigma}_{k+1} = \quad & \boldsymbol{R}_{k+1} + \boldsymbol{C}_{k+1}[(\boldsymbol{A}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{C}_k)\boldsymbol{P}_{k|k}\,(\boldsymbol{A}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{C}_k)^T \\ & + \boldsymbol{G}_k\,(\boldsymbol{Q}_k - \boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{S}_k^T)\,\boldsymbol{G}_k^T]\boldsymbol{C}_{k+1}^T \\ \boldsymbol{K}_{k+1} = \quad & [(\boldsymbol{A}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{C}_k)\,\boldsymbol{P}_{k|k}\,(\boldsymbol{A}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{C}_k)^T \\ & + \boldsymbol{G}_k\,\boldsymbol{Q}_k\,\boldsymbol{G}_k^T - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{S}_k^T\,\boldsymbol{G}_k^T]\boldsymbol{C}_{k+1}^T\,\boldsymbol{\Sigma}_{k+1}^{-1} \\ \boldsymbol{P}_{k+1|k+1} = & [\boldsymbol{I} - \boldsymbol{K}_{k+1}\,\boldsymbol{C}_{k+1}]\,[(\boldsymbol{A}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\boldsymbol{C}_k) \\ & (\boldsymbol{A}_k - \boldsymbol{G}_k\,\boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{C}_k)^T + \boldsymbol{G}_k(\boldsymbol{Q}_k - \boldsymbol{S}_k\,\boldsymbol{R}_k^{-1}\,\boldsymbol{S}_k^T)\,\boldsymbol{G}_k^T]\;;\; \boldsymbol{P}_{0|0} = \boldsymbol{P}_0 \end{aligned} \tag{11.96}$$

Kalman filter as a one-step ahead predictor:

The formulae of the predictor are obtained in the same manner as the formulae (11.91) and (11.92) of the predictor without correlation of the noises, by replacing the old matrices $\boldsymbol{A}_k$, $\boldsymbol{Q}_k$ with the new matrices $\boldsymbol{A}'_k$, $\boldsymbol{Q}'_k$ and by considering the term $\boldsymbol{G}'_k y_k$. The intermediate calculations are cumbersome; they use, in particular, the formulae of $\boldsymbol{\Sigma}_k^{-1}$ as in Eq. (11.85). The final predictor results from

$$\begin{aligned} \hat{\boldsymbol{x}}_{k+1|k} &= [\boldsymbol{A}_k - \boldsymbol{L}_k\,\boldsymbol{C}_k]\,\hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{B}_k\,\boldsymbol{u}_k + \boldsymbol{L}_k\,\boldsymbol{y}_k \\ \boldsymbol{L}_k &= [\boldsymbol{A}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{C}_k^T + \boldsymbol{G}_k\,\boldsymbol{S}_k]\,\boldsymbol{\Sigma}_k^{-1}\;;\; \boldsymbol{x}_{0|-1} = \boldsymbol{x}_0 \end{aligned} \tag{11.97}$$

with the covariance matrix of the predicted error solution of the discrete Riccati equation

$$\begin{aligned} \boldsymbol{P}_{k+1|k} = \; & \boldsymbol{A}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{A}_k^T + \boldsymbol{G}_k\,\boldsymbol{Q}_k\,\boldsymbol{G}_k^T - [\boldsymbol{G}_k\,\boldsymbol{S}_k + \boldsymbol{A}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{A}_k^T\,\boldsymbol{C}_k] \\ & [\boldsymbol{C}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{C}_k^T + \boldsymbol{R}_k]^{-1}[\boldsymbol{G}_k\,\boldsymbol{S}_k + \boldsymbol{A}_k\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{A}_k^T\,\boldsymbol{C}_k]^T\;;\; \boldsymbol{P}_{0|-1} = \boldsymbol{P}_0 \end{aligned} \tag{11.98}$$

The prediction of the output is equal to

$$\hat{y}_k = \boldsymbol{C}_k \, [q \, \boldsymbol{I} - \boldsymbol{A}_k + \boldsymbol{L}_k \, \boldsymbol{C}_k]^{-1} \, [\boldsymbol{B}_k \, u_k + \boldsymbol{L}_k \, y_k] \tag{11.99}$$

Many algorithms allow us improvement of the numerical implementation of the Kalman filter (Borne et al., 1990; Bozzo, 1983; Favier, 1982).

Innovation Representation

The innovation represents the part of $y(t)$ which cannot be predicted from the old values. The error between the real output and the predicted output, or output error prediction, is called the innovation form and is equal to

$$\begin{aligned} \tilde{\boldsymbol{y}}_{k|k-1} &= \boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1} \\ &= \boldsymbol{y}_k - \boldsymbol{C}_k \, \hat{\boldsymbol{x}}_{k|k-1} \\ &= \boldsymbol{C}_k \, (\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k-1}) + \boldsymbol{v}_k \\ &= \boldsymbol{C}_k \, \tilde{\boldsymbol{x}}_{k|k-1} + \boldsymbol{v}_k \end{aligned} \tag{11.100}$$

According to Eq. (11.91), the state representation can be written again in the innovation form

$$\begin{aligned} \hat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{A}_k \, \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{B}_k \, \boldsymbol{u}_k + \boldsymbol{A}_k \, \boldsymbol{K}_k \tilde{\boldsymbol{y}}_{k|k-1} \\ \boldsymbol{y}_k &= \boldsymbol{C}_k \, \hat{\boldsymbol{x}}_{k|k-1} + \tilde{\boldsymbol{y}}_{k|k-1} \end{aligned} \tag{11.101}$$

Noting that $e(t) = \tilde{\boldsymbol{y}}_{k|k-1}$, this model can be brought closer to the transfer function model (11.4) by setting

$$\begin{aligned} y(t) &= \boldsymbol{G}(q, \theta) \, u(t) + \boldsymbol{H}(q, \theta) \, e(t) \\ \boldsymbol{G}(q, \theta) &= \boldsymbol{C}_k \, [q\boldsymbol{I} - \boldsymbol{A}_k]^{-1} \, \boldsymbol{B}_k \\ \boldsymbol{H}(q, \theta) &= \boldsymbol{C}_k \, [q\boldsymbol{I} - \boldsymbol{A}_k]^{-1} \, \boldsymbol{A}_k \, \boldsymbol{K}_k + \boldsymbol{I} \end{aligned} \tag{11.102}$$

The covariance of the innovation form $\tilde{\boldsymbol{y}}_{k|k}$ is equal to

$$\begin{aligned} \boldsymbol{U}_{k|k-1} &= \mathrm{E}\left[\tilde{\boldsymbol{y}}_{k|k-1} \tilde{\boldsymbol{y}}_{k|k-1}^T\right] \\ &= \mathrm{E}\left\{[\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}][\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}]^T\right\} \\ &= \mathrm{E}\left\{[\boldsymbol{C}_k \, \tilde{\boldsymbol{x}}_{k|k-1} + \boldsymbol{v}_k][\boldsymbol{C}_k \, \tilde{\boldsymbol{x}}_{k|k-1} + \boldsymbol{v}_k]^T\right\} \\ &= \boldsymbol{C}_k \, \boldsymbol{P}_{k|k-1} \, \boldsymbol{C}_k^T + R_k \end{aligned} \tag{11.103}$$

This matrix gives information on abnormal phenomena which may occur during observation. In effect, if a component of the innovation form $\tilde{\boldsymbol{y}}_{i,k|k-1}$ becomes larger than n times the standard deviation $\sigma_{\tilde{y},i,k}$ deduced from the matrix $\boldsymbol{U}_{k|k-1}$, this may indicate a sensor failure, a model change, etc.

Example 11.1: **Parameterization of the Innovation Form**

The reasoning takes place in an example. Consider an ARMAX model for which $n_a = n_b = n_c = 2$; this model is written as

$$\begin{aligned} y(t) + a_1 \, y(t-1) + a_2 \, y(t-2) = \; & b_1 \, u(t-1) + b_2 \, u(t-2) \\ & + e(t) + c_1 \, e(t-1) + c_2 \, e(t-2) \end{aligned} \tag{11.104}$$

To transform it into the state representation form, we set the estimator of the state vector

$$\hat{\boldsymbol{x}}^T(t) = [\hat{x}_1(t) \quad \hat{x}_2(t)] \tag{11.105}$$

such that its first component verifies

$$\hat{x}_1(t) = y(t) - e(t) \tag{11.106}$$

For this reason

$$\begin{array}{rcl} \hat{x}_1(t) & = & -a_1\, y(t-1) - a_2\, y(t-2) + b_1\, u(t-1) + b_2\, u(t-2) \\ & & +c_1\, e(t-1) + c_2\, e(t-2) \\ & = & -a_1\, \hat{x}_1(t-1) - a_2\, y(t-2) + b_1\, u(t-1) + b_2\, u(t-2) \\ & & +(c_1 - a_1)\, e(t-1) + c_2\, e(t-2) \end{array} \tag{11.107}$$

We then set

$$\hat{x}_1(t) = -a_1\, \hat{x}_1(t-1) + b_1\, u(t-1) + (c_1 - a_1)\, e(t-1) + \hat{x}_2(t-1) \tag{11.108}$$

giving the second component of the state vector

$$\begin{array}{rcl} \hat{x}_2(t) & = & -a_2\, y(t-1) + b_2\, u(t-1) + c_2\, e(t-1) \\ & = & -a_2\, \hat{x}_1(t-1) + b_2\, u(t-1) \\ & & +(c_2 - a_2)\, e(t-1) \end{array} \tag{11.109}$$

In summary, the system of equations from which we draw the state representation is

$$\begin{array}{lcl} \hat{x}_1(t+1) & = & -a_1\, \hat{x}_1(t) + b_1\, u(t) + (c_1 - a_1)\, e(t) + \hat{x}_2(t) \\ \hat{x}_2(t+1) & = & -a_2\, \hat{x}_1(t) + b_2\, u(t) + (c_2 - a_2)\, e(t) \\ \hat{x}_1(t) & = & y(t) - e(t) \end{array} \tag{11.110}$$

giving the matrices (under the canonical observer form)

$$\boldsymbol{A}(\theta) = \begin{bmatrix} -a_1 & 1 \\ -a_2 & 0 \end{bmatrix},\ \boldsymbol{B}(\theta) = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},\ \boldsymbol{K}(\theta) = \begin{bmatrix} c_1 - a_1 \\ c_2 - a_2 \end{bmatrix},\ \boldsymbol{C}(\theta) = \begin{bmatrix} 1 & 0 \end{bmatrix} \tag{11.111}$$

and the parameter vector

$$\theta = [a_1\,,\, a_2\,,\, b_1\,,\, b_2\,,\, k_1\,,\, k_2]^T \tag{11.112}$$

with the equality

$$k_i = c_i - a_i \quad \forall i \tag{11.113}$$

Implementation of Discrete Linear Kalman Filter for State Estimation

Very often, the discrete Kalman filter is used for state estimation in the linear domain as a linear Kalman filter or in the nonlinear domain as an extended Kalman filter. A practical and simple implementation (Brown and Hwang, 1997) of the linear Kalman filter for state estimation, which is often used, is the following

- Prediction stage:

$$\begin{array}{lll} \hat{\boldsymbol{x}}_{k+1|k} & = \boldsymbol{A}_k\,\hat{\boldsymbol{x}}_{k|k} + \boldsymbol{B}_k\,\boldsymbol{u}_k \quad ; & \hat{\boldsymbol{x}}_{0|0} = \hat{\boldsymbol{x}}_0 \\ \boldsymbol{P}_{k+1|k} & = \boldsymbol{A}_k\,\boldsymbol{P}_{k|k}\,\boldsymbol{A}_k^T + \boldsymbol{G}_k\boldsymbol{Q}_k\boldsymbol{G}_k^T \quad ; & \hat{\boldsymbol{P}}_{0|0} = \hat{\boldsymbol{P}}_0 \end{array} \tag{11.114}$$

- Optimal gain calculation:

$$\boldsymbol{K}_{k+1} = \boldsymbol{P}_{k+1|k}\boldsymbol{C}_{k+1}^T[\boldsymbol{C}_{k+1}\boldsymbol{P}_{k+1|k}\boldsymbol{C}_{k+1}^T + \boldsymbol{R}_{k+1}]^{-1} \tag{11.115}$$

- Correction stage:

$$\begin{array}{ll} \hat{\boldsymbol{x}}_{k+1|k+1} & = \hat{\boldsymbol{x}}_{k+1|k} + \boldsymbol{K}_{k+1}\,(\boldsymbol{y}_{k+1} - \boldsymbol{C}_{k+1}\hat{\boldsymbol{x}}_{k+1|k}) \\ \boldsymbol{P}_{k+1|k+1} & = (\boldsymbol{I} - \boldsymbol{K}_{k+1}\boldsymbol{C}_{k+1})\boldsymbol{P}_{k+1|k} \end{array} \tag{11.116}$$

  $\boldsymbol{y}_{k+1}$ are the available measurements. In the case of a stationary model, the matrices $\boldsymbol{A}_k$, $\boldsymbol{B}_k$, $\boldsymbol{C}_k$ are constant. In the case of a nonlinear model, they are taken as the Jacobian matrices of this model at the operating point $(\boldsymbol{x}_k, \boldsymbol{u}_k)$ and thus are varying matrices.

Example 11.2: **State estimation for a pilot plant evaporator**

(Hamilton et al., 1973) describes in detail the use of a linear Kalman filter to estimate the states of a pilot plant evaporator whose model is given by (**?**).

Smoothing Kalman Filter

A rather frequent application of the Kalman filter is the smoothing filter. This problem, in particular, occurs in the case of measurements realized in a process with different sampling periods and delays (Lucena et al., 1995). Having the measurements in an interval $[0, N]$, we wish to estimate the state $\boldsymbol{x}$ at an intermediate instant i between the initial instant 0 and the final instant N. Compared to i, we thus have past information denoted by $\hat{\boldsymbol{x}}^f$ (forward) and future information $\hat{\boldsymbol{x}}^b$ (backward) (Borne et al., 1990). The past estimations $\hat{\boldsymbol{x}}^f$ can be calculated from Eqs. (11.90) between 0 and i. For future estimations $\hat{\boldsymbol{x}}^b$, we consider the system evolving in the opposite time direction for k between N and i, thus by setting $k' = N - k$

$$\left\{ \begin{array}{ll} \boldsymbol{x}_{k'+1} = & -\boldsymbol{A}_{k'}\,\boldsymbol{x}_{k'} - \boldsymbol{B}_{k'}\,\boldsymbol{u}_{k'} - \boldsymbol{G}_{k'}\,\boldsymbol{w}_{k'} \\ \boldsymbol{y}_{k'} = & \boldsymbol{C}_{k'}\,\boldsymbol{x}_{k'} + \boldsymbol{v}_{k'} \end{array} \right. \tag{11.117}$$

in which we must apply Eqs. (11.90) for k' varying from 0 to $N - i$. In the filter equations, it suffices to change $\boldsymbol{A}_k$ to $-\boldsymbol{A}_{k'}$, similarly for $\boldsymbol{B}_k$ and $\boldsymbol{G}_k$. The matrices $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$ and $\boldsymbol{C}_k$ are unchanged. The initial conditions for $k' = 0$ correspond to the final conditions for k.

The estimation $\boldsymbol{x}_i$ is the optimal weighting of the estimations obtained in the forward way denoted by f and the backward way denoted by b

$$\boldsymbol{x}_i = [\boldsymbol{P_i^f}^{-1} + \boldsymbol{P_i^b}^{-1}]^{-1}[\boldsymbol{P_i^f}^{-1}\,\boldsymbol{x}_i^f + \boldsymbol{P_i^b}^{-1}\,\boldsymbol{x}_i^b] \tag{11.118}$$

where $\boldsymbol{P_i^f}$ is the covariance matrix of the estimation error obtained by the forward formula at time i (similarly for the backward b).

Stationary Kalman Filter

The calculation of the state estimations is made easier in the case where the stationary form of the Kalman filter is considered. Moreover, the calculation will be faster. Then, the filter is not optimal anymore, but is only a sub-optimal filter. However, if the first estimations are different, the subsequent estimations will converge towards the same values as for the iterative filter.

The formulae of the stationary Kalman filter are obtained by considering the derivative of the error covariance matrix to be zero in the continuous case: $\dot{\boldsymbol{P}}(t) = 0$, and the variation of this matrix to be zero in the discrete case: $\boldsymbol{P}_{k+1} = \boldsymbol{P}_k$.

In the discrete case, we must consider the discrete Riccati equation (11.92) which becomes

$$\boldsymbol{P} = \boldsymbol{A}\,\boldsymbol{P}\,\boldsymbol{A}^T + \boldsymbol{G}\,\boldsymbol{Q}\,\boldsymbol{G}^T - \boldsymbol{A}\,\boldsymbol{P}\,\boldsymbol{C}^T\,[\boldsymbol{C}\,\boldsymbol{P}\,\boldsymbol{C}^T + \boldsymbol{R}]^{-1}\boldsymbol{C}\,\boldsymbol{P}^T\,\boldsymbol{A}^T \tag{11.119}$$

and the Kalman gain will be constant and calculated according to Eq. (11.91), transformed into

$$\boldsymbol{K} = \boldsymbol{P}\,\boldsymbol{C}^T\,[\boldsymbol{C}\,\boldsymbol{P}\,\boldsymbol{C}^T + \boldsymbol{R}]^{-1} \tag{11.120}$$

11.2 Models of Time-Varying Linear Systems

It suffices to express that the model parameters are no more constant, but are time-dependent, i.e.

$$y(t) = \sum_{k=1}^{\infty} g_t(k)\,u(t-k) + v(t) \tag{11.121}$$

In this case, the time-varying parameters $g_t(k)$ are, in fact, the impulse response at time t and thus appear as a weighting function of the input.

In the case of the state representation, the matrices containing the parameters simply depend on time

$$\begin{array}{lcl} \boldsymbol{x}(t+1,\theta) & = & \boldsymbol{A}_t(\theta)\,\boldsymbol{x}(t,\theta) + \boldsymbol{B}_t(\theta)\,u(t) + \boldsymbol{K}_t(\theta)\,e(t) \\ y(t) & = & \boldsymbol{C}_t(\theta)\,\boldsymbol{x}(t,\theta) + e(t) \end{array} \tag{11.122}$$

The Kalman filter can be applied to linear systems with time-varying parameters.

11.3 Linearization of Nonlinear Time-Varying Models

Consider a model of a nonlinear system, expressed in the state space

$$\begin{array}{lcl} x(t+1) & = & f[x(t),u(t)] + l[x(t),u(t)]\, w(t) \\ y(t) & = & h[x(t)] + m[x(t),u(t)]\, v(t) \end{array} \tag{11.123}$$

where $w(t)$ and $v(t)$ are white noises, respectively, of the model and the measurement, l and m are scalar functions. By assuming that the nominal regime corresponds to a sequence of inputs $u^*(t)$ and a corresponding trajectory $x^*(t)$, the system can be linearized in the neighbourhood of the nominal trajectory by a Taylor first-order expansion by neglecting the higher-order terms

$$\begin{array}{lcl} \tilde{x}(t+1) & = & A(t)\,\tilde{x}(t) + B(t)\,\tilde{u}(t) + w'(t) \\ \tilde{y}(t) & = & C(t)\,\tilde{x}(t) + v'(t) \end{array} \tag{11.124}$$

with the deviations with respect to the trajectory

$$\tilde{x}(t) = x(t) - x^*(t)\,,\ \tilde{y}(t) = y(t) - h[x^*(t)]\,,\ \tilde{u}(t) = u(t) - u^*(t) \tag{11.125}$$

and

$$A(t) = \frac{\partial}{\partial x} f(x,u)|_{x^*(t),u^*(t)}\,;\ B(t) = \frac{\partial}{\partial u} f(x,u)|_{x^*(t),u^*(t)}\,;\ C(t) = \frac{\partial}{\partial x} h(x)|_{x^*(t)} \tag{11.126}$$

Given the approximation around the reference, the noises of the linearized model can be considered as white noises.

11.4 Principles of Parametric Estimation

The principle of parametric estimation is to find a parameter vector that is able to make the prediction errors as low as possible. The considered models are linear with respect to the parameters to be identified.

11.4.1 Minimization of Prediction Errors

The sequence of the prediction errors $\{\epsilon(t,\theta)\}_{1\le t\le N}$ can be considered as a vector for which it is necessary to define a norm V_N that is to be minimized. The norm depends on the parameter set θ and on the data set Z^N.

For example, use a linear stable filter $L(q)$ such that

$$\epsilon_f(t,\theta) = L(q)\,\epsilon(t,\theta)\,, \quad 1 \le t \le N \tag{11.127}$$

We can then use a norm, which will be commonly in the form

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} l(\epsilon_f(t,\theta)) \tag{11.128}$$

where l is a scalar function.

We will then deduce the estimation $\hat{\theta}$ of the parameter vector as the vector minimizing this norm V_N

$$\hat{\theta}(Z^N) = \arg\left\{\min_{\theta} V_N(\theta, Z^N)\right\} \tag{11.129}$$

This general technique of minimization constitutes the PEM (prediction error methods) family (Ljung, 1987).

The aim of the filter $L(q)$ is to realize a weighting function acting on the frequencies, thus on the noise.

Consider the classical case where the norm is a quadratic criterion (which facilitates derivation)

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{2}\epsilon^2(t,\theta) \tag{11.130}$$

On the other hand, suppose that the process model, expressed as a transfer functions, is the following

$$y(t) = G(q,\theta)\,u(t) + H(q,\theta)\,e(t) \tag{11.131}$$

In this case, the prediction error is equal to

$$\epsilon(t,\theta) = H^{-1}(q,\theta)\ [y(t) - G(q,\theta)\,u(t)] \tag{11.132}$$

According to Ljung (1987), the discrete Fourier transform (DFT) of the sequence $\{\epsilon(t,\theta)\}$ is equal to

$$E_N(2\pi k/N,\theta) = \frac{1}{\sqrt{N}}\sum_{t=1}^{N}\epsilon(t,\theta)\exp(-i2\pi kt/N)\ ,\ k = 0,\ldots,N-1 \tag{11.133}$$

The chosen norm corresponds to the signal energy. Thus, it is possible to apply the Parseval-Plancherel relation to this signal. We thus get in the frequency domain

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{k=O}^{N-1}\frac{1}{2}|E_N(2\pi k/N,\theta)|^2 \tag{11.134}$$

$y(t)$ can be separated into two parts, one related to the input, the other one related to the noise; thus

$$w(t,\theta) = G(q,\theta)\,u(t)\ ,\quad s(t,\theta) = H(q,\theta)\,e(t) \tag{11.135}$$

Their DFT are respectively equal to

$$W_N(\omega,\theta) = G(\exp(i\omega),\theta)\,U_N(\omega) + R_{N,1}(\omega)\ ,\quad \text{with: } |R_{N,1}(\omega)| \le \frac{C_1}{\sqrt{N}} \tag{11.136}$$

assuming that the transfer function G is stable and the input $u(t)$ is bounded, and

$$S_N(\omega,\theta) = Y_N(\omega,\theta) - G(\exp(i\omega),\theta)\, U_N(\omega) - R_{N,1}(\omega) \tag{11.137}$$

As we can express the prediction error by

$$\epsilon(t,\theta) = H^{-1}(q,\theta)\, s(t,\theta) \tag{11.138}$$

the DFT of the prediction error can be expressed as

$$E_N(\omega,\theta) = H^{-1}(\exp(i\omega),\theta)\, S_N(\omega) + R_{N,2}(\omega)\ , \quad \text{with: } |R_{N,\epsilon}(\omega)| \leq \frac{C_2}{\sqrt{N}} \tag{11.139}$$

assuming that the transfer function H is stable and the signal $s(t)$ is bounded. The norm to be minimized results

$$\begin{aligned} V_N(\theta, Z^N) = \quad & \frac{1}{N}\sum_{k=1}^{N} \frac{1}{2}\, |H^{-1}(\exp(i2\pi k/N),\theta)|^2\, |Y_N(2\pi k/N),\theta) - \\ & G(2\pi k/N),\theta)\, U_N(2\pi k/N),\theta)|^2 + R_{N,3}\ , \text{ with: } |R_{N,3}(\omega)| \leq \tfrac{C_3}{\sqrt{N}} \end{aligned} \tag{11.140}$$

Because of the frequency form of this expression, the identification methods by prediction error are closely related to the methods of spectral analysis (Ljung, 1987).

11.4.2 Linear Regression and Least Squares

The models of linear regression provide the output prediction

$$\hat{y}(t|\theta) = \phi^T(t)\, \theta \tag{11.141}$$

where ϕ is the observation vector and θ is the parameter vector.

Estimation of the Parameter Vector

The prediction error is equal to

$$\epsilon(t,\theta) = y(t) - \phi^T(t)\, \theta \tag{11.142}$$

Retaining the previous notations, the filter $L(q)$ is taken to be equal to 1 and the chosen norm to be minimized is

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \frac{1}{2}\, \left[y(t) - \phi^T(t)\, \theta\right]^2 \tag{11.143}$$

The ratio $1/2$ is only introduced for derivation reasons and the ratio $1/N$ could be omitted without modifying the results. This norm is a quadratic function with respect to θ. It suffices to express that the condition of minimum of the

norm implies that the gradient vector is zero at the extremum, thus that the partial derivatives with respect to θ are zero. We deduce the estimator of the parameter vector according to the least-squares criterion

$$\hat{\theta}^{LS} = \left[\frac{1}{N}\sum_{t=1}^{N}\phi(t)\,\phi^T(t)\right]^{-1}\;\frac{1}{N}\sum_{t=1}^{N}\phi(t)\,y(t) = \left[\sum_{t=1}^{N}\phi(t)\,\phi^T(t)\right]^{-1}\;\sum_{t=1}^{N}\phi(t)\,y(t) \tag{11.144}$$

provided that the inverse of the following matrix exists

$$R(N) = \frac{1}{N}\sum_{t=1}^{N}\phi(t)\,\phi^T(t) \tag{11.145}$$

Recall that for the ARX model, the observation vector $\phi(t)$ was equal to

$$\phi(t) = \left[-y(t-1)\,,\,\ldots\,,\,-y(t-n_a)\,,\,u(t-1)\,,\,\ldots\,,\,u(t-n_b)\right]^T \tag{11.146}$$

The terms of the matrix $R(N)$ are thus

$$[R(N)]_{ij} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{t=1}^{N} y(t-i)\,y(t-j) & \text{if: } 1 \le i,j \le n_a \\ -\dfrac{1}{N}\displaystyle\sum_{t=1}^{N} u(t-i+n_a)\,y(t-j) & \text{if: } n_a < i \le n_a+n_b \text{ and } 1 \le j \le n_a \\ -\dfrac{1}{N}\displaystyle\sum_{t=1}^{N} y(t-i)\,u(t-j+n_a) & \text{if: } 1 \le i \le n_a \text{ and } n_a < j \le n_a+n_b \\ \dfrac{1}{N}\displaystyle\sum_{t=1}^{N} u(t-i+n_a)\,u(t-j+n_a) & \text{if: } n_a < i,j \le n_a+n_b \end{cases} \tag{11.147}$$

which implies that the matrix $R(N)$ is, in fact, composed of estimations of the covariances of $\{y(t)\}$ and $\{u(t)\}$ whose calculation allows us to deduce the best estimation of the parameter vector.

This multiple linear regression is often presented in matrix form

$$\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\Phi}\,\theta \tag{11.148}$$

where $\boldsymbol{\epsilon}$ is the vector of the prediction errors, $\boldsymbol{Y}$ is the output vector and $\boldsymbol{\Phi}$ is the matrix gathering the vectors ϕ for the set of the experiences. θ remains the parameter vector. The chosen norm is equal to

$$V_N(\theta, Z^N) = \frac{1}{2}[\boldsymbol{Y} - \boldsymbol{\Phi}\,\theta]^T[\boldsymbol{Y} - \boldsymbol{\Phi}\,\theta] \tag{11.149}$$

The condition of zero gradient gives

$$-\boldsymbol{Y}^T\,\boldsymbol{\Phi} + \theta^T\,(\boldsymbol{\Phi}^T\boldsymbol{\Phi}) = 0 \tag{11.150}$$

So that the solution θ exists, the matrix $(\mathbf{\Phi}^T\mathbf{\Phi})$ must be nonsingular. This matrix is semipositive definite by construction, thus it suffices that it is strictly positive. When the solution exists, it is equal to

$$\hat{\theta}^{LS} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\,\mathbf{\Phi}^T\,\boldsymbol{Y} \tag{11.151}$$

The drawback of this method is that it demands the inversion of the matrix $(\mathbf{\Phi}^T\mathbf{\Phi})$, which is often ill-conditioned, and thus poses numerical problems and for this reason is badly adapted to on-line use.

Estimator Properties

Let θ_0 be the "true" (exact if the model represents exactly the process) parameter vector responding to the model

$$y(t) = \phi^T(t)\,\theta_0 + e(t) \tag{11.152}$$

thus in matrix form

$$\boldsymbol{Y} = \mathbf{\Phi}^T\,\theta_0 + \boldsymbol{e} \tag{11.153}$$

Assuming that $e(t)$ is white noise (zero mean, variance λ^2), it is possible to verify (Söderström and Stoica, 1989) the following properties
a) The estimator $\hat{\theta}^{LS}$ of the parameter vector is non-biased

$$\mathrm{E}(\hat{\theta}^{LS}) = \theta_0 \tag{11.154}$$

thus the estimator $\hat{\theta}^{LS}$ of the parameter vector must tend towards the value θ_0 when the number of the observations tends towards infinity; on the other hand, it must be as close as possible to θ_0.

The estimator $\hat{\theta}^{LS}$ of the parameter vector is related to θ_0 by the relation

$$\hat{\theta}^{LS} = \theta_0 + (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\,e(t) \tag{11.155}$$

b) The covariance matrix of $\hat{\theta}^{LS}$ is equal to

$$\mathrm{cov}(\hat{\theta}^{LS}) = \lambda^2(\mathbf{\Phi}^T\mathbf{\Phi})^{-1} \tag{11.156}$$

c) A non-biased estimation of the variance λ^2 of white noise is given by

$$s^2(e) = \hat{\lambda^2} = 2\frac{V(\hat{\theta})}{N-n} \tag{11.157}$$

where n is the dimension of the parameter vector θ.

From these properties, it is possible to deduce the variance of each parameter θ_i as

$$s^2(\theta_i) = \hat{\sigma^2}(\theta_i) = \mathrm{cov}_{ii}\;s^2(e) \tag{11.158}$$

where cov_{ij} is the current element of the matrix $(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$, and also the confidence interval of θ_i at the significance level α according to the Student t law

$$\hat{\theta}_i - t_{\alpha/2,N-n}s(\theta_i) < \theta_i < \hat{\theta}_i + t_{\alpha/2,N-n}s(\theta_i) \tag{11.159}$$

Weighted Least Squares

In the case of weighted least squares, the criterion simply contains a weight for the measurements which can depend on the time

$$V_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^{N} \alpha(t) \left[y(t) - \phi^T(t)\, \theta \right]^2 \tag{11.160}$$

By setting

$$\beta(N,t) = \frac{1}{N} \alpha(t) \tag{11.161}$$

we obtain

$$\hat{\theta}^{LS} = \left[\sum_{t=1}^{N} \beta(N,t) \phi(t)\, \phi^T(t) \right]^{-1} \sum_{t=1}^{N} \beta(N,t) \phi(t)\, y(t) \tag{11.162}$$

Case of a Coloured Noise

When the noise $v(t)$ is coloured (by opposition to white noise), the least-squares estimator does not converge towards the true value θ_0. In this case, Ljung (1987) introduces a linear filter describing the noise in the form

$$v(t) = \frac{1}{D(q)}\, e(t) \tag{11.163}$$

where $e(t)$ is white noise. The model becomes

$$A(q)\, y(t) = B(q)\, u(t) + \frac{1}{D(q)}\, e(t) \Longleftrightarrow A(q)\, D(q)\, y(t) = B(q)\, D(q)\, u(t) + e(t) \tag{11.164}$$

to which the least squares can be applied, which allows us to estimate the parameters of polynomials AD and BD (method of repeated least squares).

The best non-biased estimator (Söderström and Stoica, 1989) is equal to

$$\hat{\theta} = [\boldsymbol{R}^{-1}\, \boldsymbol{\Phi} (\boldsymbol{\Phi}^T\, \boldsymbol{R}^{-1}\, \boldsymbol{\Phi})^{-1}]^T\, \boldsymbol{Y} = [\boldsymbol{\Phi}^T\, \boldsymbol{R}^{-1}\, \boldsymbol{\Phi}]^{-1}\, \boldsymbol{\Phi}^T\, \boldsymbol{R}^{-1}\, \boldsymbol{Y} \tag{11.165}$$

with

$$\mathrm{E}(\boldsymbol{v}\boldsymbol{v}^T) = \boldsymbol{R} \tag{11.166}$$

11.4.3 Maximum Likelihood Method

Principle of Maximum Likelihood

The maximum likelihood is first exposed in a general manner following Kendall and Stuart (1979). Let $q_1, \ldots, q_n$ be a set of exclusive propositions and H the available information. Let p be a subsequent proposition which is also exclusive. From the probability properties, it results that

$$\mathcal{P}(q_i, p|H) = \mathcal{P}(q_i|H)\, \mathcal{P}(p|q_i, H) \tag{11.167}$$

and

$$\mathcal{P}(q_i|p,H) = \frac{\mathcal{P}(q_i|H)\,\mathcal{P}(p|q_i,H)}{\mathcal{P}(p|H)} \tag{11.168}$$

and by summing the set of propositions

$$\sum_i \frac{\mathcal{P}(q_i|H)\,\mathcal{P}(p|q_i,H)}{\mathcal{P}(p|H)} = 1 \tag{11.169}$$

from which the Bayes theorem is deduced

$$\mathcal{P}(q_i|p,H) = \frac{\mathcal{P}(q_i|H)\mathcal{P}(p|q_i,H)}{\sum_i[\mathcal{P}(q_i|H)\mathcal{P}(p|q_i,H)]} = \frac{\mathcal{P}(q_i,p|H)}{\sum_i \mathcal{P}(q_i,p|H)} \tag{11.170}$$

which gives the probability of a proposition q_i when p is known. The variables $\mathcal{P}(q_i|H)$ are called prior probabilities, the variables $\mathcal{P}(q_i|p,H)$ posterior probabilities and $\mathcal{P}(p|q_i,H)$ the likelihood function (or simply likelihood), which is often denoted by $\mathcal{L}(p|q_i,H)$. The posterior probability varies as the product of the prior probability and the likelihood

$$\mathcal{P}(q_i|p,H) \propto \mathcal{P}(q_i|H)\,\mathcal{L}(p|q_i,|H) \tag{11.171}$$

The principle of the maximum likelihood stipulates that when we are faced with a choice of hypotheses q_i, we choose the one which maximizes $\mathcal{L}$.

Now, place ourselves in the context of parametric identification (Ljung, 1987). The observations $y(1),\ldots,y(N)$ may not be perfectly reliable; with each observation of the observation vector $\boldsymbol{y} = \{y(1),\ldots,y(N)\}$, we can associate a probability density function $p(y|\theta)$, and for the set $\boldsymbol{y}$ of the observations considered as independent the joint probability density function $\mathcal{L}(\boldsymbol{y}|\theta)$. Let $\boldsymbol{y}^e$ be the observed value (e = experimental) of vector $\boldsymbol{y}$; to this vector corresponds an estimation $\hat{\theta}^e$ of the parameter vector. The probability that $\boldsymbol{y}$ can take the value $\boldsymbol{y}^e$ is proportional to the likelihood function $\mathcal{L}(\boldsymbol{y}^e|\theta)$. In these conditions, given the observation vector, the estimator of the parameter vector θ according to the maximum likelihood method is equal to

$$\hat{\theta}^{ML}(\boldsymbol{y}^e) = \arg\left\{\max_\theta \mathcal{L}(\boldsymbol{y}^e|\theta)\right\} \tag{11.172}$$

The maximum with respect to parameter vector θ is thus searched in order to make the vector $\boldsymbol{y}$ as close as possible to the fixed value $\boldsymbol{y}^e$.

Example 11.3: **Estimation by Maximum Likelihood**

For a set $\boldsymbol{x}^e$ of N variables x_i distributed according to the normal law (unknown mean μ_0, known variances σ_i^2), the likelihood function is equal to

$$\mathcal{L}(\boldsymbol{x}^e|\mu) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \tag{11.173}$$

The estimator of the mean according to the maximum likelihood gives

$$\begin{aligned}\hat{\mu}^{ML}(\boldsymbol{x}^e) &= \arg\{\max_\mu \ \mathcal{L}(\boldsymbol{x}^e|\mu)\} \\ &= \frac{1}{\sum_{i=1}^N (1/\sigma_i^2)} \sum_{i=1}^N \frac{x_i}{\sigma_i^2}\end{aligned} \tag{11.174}$$

We notice that when the variances are equal, this estimator gives the same result as the sample mean

$$\hat{\mu}(\boldsymbol{x}^e) = \frac{1}{N}\sum_{i=1}^N x_i \tag{11.175}$$

We can estimate the quality of an estimator $\hat{\theta}$ with respect to its true value θ_0 (in general unknown) by the expectation of the covariance matrix

$$P = \mathrm{E}\left[\hat{\theta}(y^e) - \theta_0\right]\left[\hat{\theta}(y^e) - \theta_0\right]^T \tag{11.176}$$

which we try to make as small as possible. But the Cramer-Rao inequality (Kendall and Stuart, 1979; Ljung, 1987) stipulates that there exists a limit lower than P

$$\mathrm{E}\left[\hat{\theta}(y^e) - \theta_0\right]\left[\hat{\theta}(y^e) - \theta_0\right]^T > \mathcal{I}^{-1}(\theta) \tag{11.177}$$

where $\mathcal{I}(\theta)$ is the Fisher information matrix defined by

$$\mathcal{I}(\theta) = \mathrm{E}\left[\frac{d}{d\theta}\log\mathcal{L}(\boldsymbol{y}^e|\theta)\right]\left[\frac{d}{d\theta}\log\mathcal{L}(\boldsymbol{y}^e|\theta)\right]^T_{|\theta=\theta_0} \tag{11.178}$$

In fact, the estimator $\hat{\theta}_{ML}$ converges asymptotically towards a normal distribution of mean θ_0 and of covariance matrix $\mathcal{I}^{-1}(\theta_0)$ (Kendall and Stuart, 1979).

Estimator Determination

Now consider an input-output probabilistic model; let Z^t be the set of input and output data until time t. The predictor is represented by the model

$$\hat{y}(t|t-1) = g(t, Z^{t-1}, \theta) \tag{11.179}$$

The probability density function $p_y(y(t) = y_0|Z^{t-1}, \theta)$ corresponds to the probability that $y(t)$ is equal to some value y_0, given Z^{t-1} and θ, and is simply denoted by $p_y(y(t)|Z^{t-1}, \theta)$. With the previous model is associated the probability density function $p_e(\epsilon(t) = \epsilon_0|Z^{t-1}, \theta)$ of the prediction error $\epsilon = y(t) - \hat{y}(t|t-1)$. The model is then in the form

$$y(t) = g(t, Z^{t-1}, \theta) + \epsilon(t, \theta) \tag{11.180}$$

In these conditions, the likelihood function is equal to

$$\mathcal{L}_y(y^e|\theta) = \prod_{t=1}^{N} p_e(y(t) - g(t, Z^{t-1}; \theta)|\theta) = \prod_{t=1}^{N} p_e(\epsilon(t, \theta)|\theta) \tag{11.181}$$

The maximization of this monotonous function is equivalent to that of its logarithm; we will thus maximize

$$\frac{1}{N} \log \mathcal{L}_y(y^e|\theta) = \frac{1}{N} \sum_{t=1}^{N} \log p_e(\epsilon(t, \theta)|\theta) \tag{11.182}$$

The following function, related to the logarithm of the likelihood, and also related to the notion of information entropy and the information matrix, is frequently introduced

$$l(\epsilon, \theta, t) = -\log p_e(\epsilon(t, \theta)|\theta) \tag{11.183}$$

The estimator is deduced according to the method of maximum likelihood

$$\hat{\theta}^{ML}(y^N) = \arg \left\{ \min_{\theta} \frac{1}{N} \sum_{t=1}^{N} l(\epsilon, \theta, t) \right\} \tag{11.184}$$

This method of maximum likelihood may be intricate and may necessitate resorting to approximations related to the difficulty of expressing the predictor, often in the form of a Kalman predictor.

In the case where the prediction errors are distributed in a Gaussian way (zero mean, covariance λ^2 independent of t) , we obtain

$$l(\epsilon, \theta, t) = \text{const} + \frac{1}{2} \log \lambda + \frac{1}{2} \frac{\epsilon^2}{\lambda^2} \tag{11.185}$$

11.4.4 Correlation of Prediction Errors with Past Data

A good model is characterized by the fact that the prediction errors that it provides are independent of the past data. Let Z^{t-1} be the set of past data at time $t-1$; the predictor $\hat{y}(t|\theta)$ is said to be ideal if the prediction error $\epsilon(t, \theta)$ is independent of the past data Z^{t-1}. To check it, we must take a vector sequence $\{\zeta(t)\}$ and verify that it is uncorrelated with any linear transformation of $\{\epsilon(t, \theta)\}$. In practice (Ljung, 1987), we choose a linear filter $L(q)$, which we make act on $\{\epsilon(t, \theta)\}$

$$\epsilon_F(t, \theta) = L(q)\, \epsilon(t, \theta) \tag{11.186}$$

Another sequence of correlation vectors denoted by $\{\zeta(t, Z^{t-1}, \theta)\}$ or simply $\{\zeta(t, \theta)\}$, constructed from the past data and from θ, is chosen, as well as a function $\alpha(\epsilon)$. At last, we calculate the estimation of θ such that

$$\frac{1}{N} \sum_{t=1}^{N} \zeta(t, \theta)\, \alpha(\epsilon_F(t, \theta)) = 0 \tag{11.187}$$

This estimator is the best that can be found from the past data.
Remark:
In the case of pseudolinear regressions, the prediction models are in the form

$$\hat{y}(t|\theta) = \phi^T(t,\theta)\,\theta \tag{11.188}$$

We choose $\zeta(t,\theta) = \phi(t,\theta)$ and $\alpha(\epsilon) = \epsilon$, so that the estimator of the pseudolinear regression is the solution of

$$\frac{1}{N}\sum_{t=1}^{N} \phi(t,\theta)\left[y(t) - \phi^T(t,\theta)\,\theta\right] = 0 \tag{11.189}$$

This pseudolinear designation comes from the fact that if $\phi^T(t,\theta)$ depends effectively on θ, the predictor $\hat{y}(t|\theta)$ is not linear with respect to θ. The linear regression according to the least squares is a particular case of the pseudolinear regression.

11.4.5 Instrumental Variable Method

The instrumental variable method (Söderström and Stoica, 1983) is a variant of the least-squares method which enables to avoid the bias of the estimated parameter vector when the disturbance is not white noise. Moreover, the complexity of this identification method is moderate.

Method Principle

Assuming that the regression model is linear as

$$\hat{y}(t|\theta) = \phi^T(t)\,\theta \tag{11.190}$$

an estimator of the parameter vector according to the least squares is obtained by minimization of the following criterion

$$\begin{aligned}
\hat{\theta} = \quad & \arg\left\{ \min_\theta \frac{1}{N}\sum_{t=1}^{N} \left[y^T(t) - \theta^T\,\phi(t)\right]\,\left[y(t) - \phi^T(t)\,\theta\right] \right\} \Longrightarrow \\
& \frac{1}{N}\sum_{t=1}^{N} \phi(t)\,\left[y(t) - \phi^T(t)\,\theta\right] = 0 \Longrightarrow \\
\hat{\theta} = \quad & \left[\frac{1}{N}\sum_{t=1}^{N} \phi(t)\phi^T(t)\right]^{-1} \left[\frac{1}{N}\sum_{t=1}^{N} \phi(t)\,y(t)\right]
\end{aligned} \tag{11.191}$$

From the viewpoint of the correlation of the prediction errors with the past values, this would correspond to $L(q) = 1$ and $\zeta(t,\theta) = \phi(t)$. If the system is, in fact, described by a model, including a disturbance $v(t)$, such that

$$y(t) = \phi^T(t)\,\theta_0 + v(t) \qquad \text{with: } v(t) = H(q^{-1})\,e(t) \tag{11.192}$$

the observation vector $\phi(t)$ is then correlated with $v(t)$ and the estimator does not tend anymore towards the true value θ_0 (the estimation is not consistent). From Eqs. (11.191) and (11.192), we deduce

$$\hat{\theta} = \theta_0 + \left[\frac{1}{N}\sum_{t=1}^{N}\phi(t)\phi^T(t)\right]^{-1}\left[\frac{1}{N}\sum_{t=1}^{N}\phi(t)\,v(t)\right] \tag{11.193}$$

The condition that $\hat{\theta}$ converges towards the true value θ_0 is that the disturbances is not correlated with the outputs

$$\mathrm{E}[\phi(t)\,v(t)] = 0 \tag{11.194}$$

The instrumental variable method is designed so as to remedy to this possible bias of the estimator θ when the least squares are used.

Introduce a vector $\zeta(t)$, whose components are called the instruments or instrumental variables, and are such that the instrumental variable vector $\zeta(t)$ is not correlated with $v(t)$. The estimator is then given by the solution of the following system

$$\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\;\left[y(t) - \phi^T(t)\,\theta\right] = \frac{1}{N}\sum_{t=1}^{N}\zeta(t,\theta)\,\epsilon(t) = 0 \tag{11.195}$$

Provided that the following inverse matrix exists, the estimator is then given by

$$\theta^{VI} = \left[\frac{1}{N}\sum_{t=1}^{N}\zeta(t,\theta)\,\phi^T(t)\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\zeta(t,\theta)\,y(t) \tag{11.196}$$

The conditions that the instrumental variable vector must fulfil are

$$\begin{cases} \mathrm{E}\left[\zeta(t,\theta)\,\phi^T(t)\right] & \text{is nonsingular} \\ \mathrm{E}\left[\zeta(t,\theta)\,v(t)\right] = 0 \end{cases} \tag{11.197}$$

The instrumental variables must be sufficiently correlated with the regression variables so that the matrix is invertible but not correlated with the noise.

The implementation of the instrumental variable method is described in Sect. 12.3.6.

Application to an ARX Model

Suppose that the used model is an ARX model such as

$$y(t) + a_1\,y(t-1) + \ldots + a_{n_a}\,y(t-n_a) = b_1\,u(t-1) + \ldots + b_{n_b}\,y(t-n_b) + v(t) \tag{11.198}$$

while the true model would be

$$y(t)+(a_1)_0\,y(t-1)+\ldots+(a_{n_a})_0\,y(t-n_a) = (b_1)_0\,u(t-1)+\ldots+(b_{n_b})_0\,y(t-n_b)+v(t) \tag{11.199}$$

In order to ensure the conditions necessary for the instrumental variables, we choose them to be equal to

$$\zeta(t) = K(q)\left[-x(t-1),\, -x(t-2),\, \ldots,\, -x(t-n_a),\, u(t-1),\, \ldots,\, u(t-n_b)\right]^T \tag{11.200}$$

where $K(q)$ is a linear filter and $x(t)$ is generated from the input $u(t)$ through a linear system according to

$$x(t) = \frac{N(q)}{D(q)}\, u(t) \tag{11.201}$$

with the polynomials of the transfer function

$$\begin{array}{rcl} N(q) & = & m_0 + m_1\, q^{-1} + \ldots + m_{n_m}\, q^{-n_m} \\ D(q) & = & 1 + n_1\, q^{-1} + \ldots + n_{n_n}\, q^{-n_n} \end{array} \tag{11.202}$$

The instrumental variables $\zeta(t)$ depend on the past inputs $u(t-1)\ldots$ through a linear filter.

This method can be used in open loop, as the inputs do not then depend on the noise $v(t)$ of the system; it will then have to be modified in closed loop.

Ideally, the filter should be equal to the transfer function of the system with $N(q) = B(q)$ and $D(q) = A(q)$ or

$$x(t,\theta) = \frac{B(q)}{A(q)}\, u(t) \tag{11.203}$$

so that the instrumental variables depend on the parameter vector θ.

References

P. Borne, G. Dauphin-Tanguy, J.P. Richard, F. Rotella, and I. Zambettakis. *Commande et Optimisation des Processus.* Technip, Paris, 1990.

C.A. Bozzo. *Le Filtrage Optimal et ses Applications aux Problèmes de Poursuite*, volume 2 of *Théorie de l'Estimation, Propriétés des Estimateurs en Temps Discret et Applications.* Lavoisier, Paris, 1983.

R.G. Brown and P.Y.C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering.* Wiley, New York, third edition, 1997.

G. Favier. *Filtrage, Modélisation et Identification de Systèmes Linéaires Stochastiques.* CNRS, Paris, 1982.

G.C. Goodwin and K.S. Sin. *Adaptive Filtering, Prediction and Control.* Prentice Hall, Englewood Cliffs, 1984.

M.S. Grewal and A.P. Andrews. *Kalman Filtering Theory and Practice Using MATLAB.* Wiley, New York, 2nd edition, 2001.

M.S. Grewal and A.P. Andrews. *Kalman Filtering: Theory and Practice*. Prentice Hall, Englewood Cliffs, NJ, 1993.

J.C. Hamilton, D.E. Seborg, and D.G. Fisher. An experimental evaluation of Kalman filtering. *AIChE J.*, 19(5):901–908, 1973.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D, J. Basic Eng.*, 82:35–45, 1960.

R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME Ser. D, J. Basic Eng.*, 83:95–108, 1961.

E.W. Kamen and J.K. Su. *Introduction to Optimal Estimation*. Springer-Verlag, London, 1999.

M. Kendall and A. Stuart. *The Advanced Theory of Statistics*. Charles Griffin, London, 1979.

I.D. Landau and A. Besançon Voda, editors. *Identification des Systèmes*. Hermès, Paris, 2001.

L. Ljung. *System Identification. Theory for the User*. Prentice Hall, Englewood Cliffs, 1987.

S. Lucena, C. Fonteix, I. Marc, and J.P. Corriou. Nonlinear control of a discontinuous bioreactor with measurement delays and state estimation. In A. Isidori, editor, *European Control Conference ECC95*, volume 4, pages 3811–3815, Rome, Italie, 1995.

R.H. Middleton and G.C. Goodwin. *Digital Control and Estimation*. Prentice Hall, Englewood Cliffs, 1990.

J.C. Radix. *Introduction au Filtrage Numérique*. Eyrolles, Paris, 1970.

T. Söderström and P. Stoica. *System Identification*. Springer-Verlag, Berlin, 1983.

T. Söderström and P. Stoica. *System Identification*. Prentice Hall, New York, 1989.

E. Walter, editor. *Identifiability of Parametric Models*. Pergamon, Oxford, 1987.

E. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data*. Communications and Control Engineering. Springer-Verlag, London, 1997.

K. Watanabe. *Adaptive Estimation and Control*. Prentice Hall, London, 1992.

P. Zarchan and H. Musoff. *Fundamentals of Kalman Filtering: a Practical Approach*, volume 190 of *AIAA Progress Series in Astronautics and Aeronautics*. AIAA, Reston, VA, 2000.

Chapter 12

Parametric Estimation Algorithms

In view of adaptive control, many algorithms of process model parameter estimation have been developed.

Recall that, according to the used hypotheses (Ljung and Söderström, 1986), the aim is to make the predictions at time t, knowing the inputs and outputs up to $t-1$ which form the observation vector denoted by $\phi(t)$, while the estimated parameter vector is denoted by $\theta(t)$. Some authors (Dugard and Landau, 1990; Landau, 1988, 1990) make the prediction at $t+1$ and denote by $\phi(t)$ the observation vector up to t, while the estimated parameter vector is denoted by $\theta(t+1)$. The difference thus bears on the adopted notation for $\phi(t)$ and the prediction instant. This will lead to subscript differences in the formulae; thus it is necessary to carefully look at the bases of the reference textbook.

12.1 Linear Regression and Least Squares

This algorithm, which is the simplest, has already been discussed in a previous chapter. It is essentially devoted to an on-line usage, i.e. to estimate the parameters of a fixed model. Here, we will note some problems concerning it.

We assume that the prediction model is

$$\hat{y}(t|\theta) = \phi^T(t)\,\theta \tag{12.1}$$

The estimator of the parameter vector minimizing the quadratic criterion of the sum of the squares of the prediction errors

$$J_N(\theta) = \frac{1}{2}\sum_{t=1}^{N}(y(t) - \hat{y}(t|\theta))^2 \tag{12.2}$$

can be written in the form

$$\hat{\theta}^{LS} = \left[\frac{1}{N} \sum_{t=1}^{N} \phi(t)\, \phi^T(t) \right]^{-1} \frac{1}{N} \sum_{t=1}^{N} \phi(t)\, y(t) \tag{12.3}$$

provided that the inverse of the matrix introduced exists. Set the matrix $\boldsymbol{R}$

$$\boldsymbol{R}(N) = \frac{1}{N} \sum_{t=1}^{N} \phi(t)\, \phi^T(t) \tag{12.4}$$

and the vector $f(N)$

$$f(N) = \frac{1}{N} \sum_{t=1}^{N} \phi(t)\, y(t) \tag{12.5}$$

which implies that the estimator of the parameter vector is the solution of the system

$$\boldsymbol{R}(N)\, \hat{\theta}^{LS} = f(N) \tag{12.6}$$

The algorithm thus defined is non-recursive.

The matrix $\boldsymbol{R}(N)$ can be badly conditioned (the ratio of the largest singular value to the smallest one is very large), in particular when the inputs are not sufficiently exciting. In this case, it is better to seek a matrix $\boldsymbol{Q}$ such that

$$\boldsymbol{Q}\, \boldsymbol{Q}^T = \boldsymbol{R}(N) \tag{12.7}$$

which we can obtain by the Householder or Gram-Schmidt transformation, or the Cholesky decomposition (Golub and Loan, 1989).

Remark 1:

(Ljung, 1987) shows that this decomposition of $\boldsymbol{R}$ improves the conditioning of the matrix. This demonstration can be done in the multivariable case (m inputs, p outputs) by setting

$$Y^T = \left[y(1)^T, \ldots, y(N)^T \right]^T \tag{12.8}$$

where $y(t)$ is the output vector of dimension p. Let

$$\boldsymbol{\Phi}^T = \left[\phi(1), \ldots, \phi(N) \right]^T \tag{12.9}$$

The criterion to be minimized is equal to

$$J_N(\theta) = |Y - \boldsymbol{\Phi}\, \theta|^2 = \sum_{t=1}^{N} (y(t) - \phi^T(t)\, \theta)^2 \tag{12.10}$$

This norm is not modified by an orthonormal transformation $\boldsymbol{P}$ (a matrix $\boldsymbol{P}$ is orthonormal if $\boldsymbol{P}\, \boldsymbol{P}^T = \boldsymbol{I}$), so that

$$J_N(\theta) = |\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\Phi}\, \theta)|^2 \tag{12.11}$$

The matrix $\boldsymbol{P}$ is chosen in order to realize a factorization QR of $\boldsymbol{\Phi}$

$$\boldsymbol{\Phi} = \boldsymbol{P}^T \begin{bmatrix} \boldsymbol{Q} \\ 0 \end{bmatrix} \tag{12.12}$$

A possibility is to take $\boldsymbol{P}$ to be equal to a product of Householder transformations such that

$$\boldsymbol{P}\,\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{M} \end{bmatrix} \tag{12.13}$$

hence

$$J_N(\theta) = \left| \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{M} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Q} \\ 0 \end{bmatrix} \theta \right|^2 = |\boldsymbol{L} - \boldsymbol{Q}\,\theta|^2 + |\boldsymbol{M}|^2 \tag{12.14}$$

We deduce that the criterion is minimal when

$$\boldsymbol{Q}\,\hat{\theta} = \boldsymbol{L} \tag{12.15}$$

giving the new value of the estimator.

We can notice that

$$\boldsymbol{Q}^T\,\boldsymbol{Q} = [\boldsymbol{P}\,\boldsymbol{\Phi}]^T\ \boldsymbol{P}\,\boldsymbol{\Phi} = \boldsymbol{\Phi}^T\,\boldsymbol{P}^T\,\boldsymbol{P}\,\boldsymbol{\Phi} = \boldsymbol{\Phi}^T\,\boldsymbol{\Phi} = \boldsymbol{R}(N) \tag{12.16}$$

Thus, the new condition number is equal to the square root of the condition number of $\boldsymbol{R}(N)$ and the new system

$$\boldsymbol{Q}\,\hat{\theta} = \boldsymbol{L} \tag{12.17}$$

is better conditioned than the old system

$$\boldsymbol{R}(N)\hat{\theta}^{LS} = f(N) \tag{12.18}$$

Remark 2:
The vector $\boldsymbol{\Phi}$ can be put in a global form

$$\boldsymbol{\Phi} = \begin{bmatrix} z(t-1) \\ \dots \\ z(t-n) \end{bmatrix} \tag{12.19}$$

where z gathers the set of the variables necessary for knowledge of ϕ, e.g. for an ARX model $z = [-y, u]$ and n is the largest shift necessary to form $\boldsymbol{\Phi}$, which is related to the structure of the model. According to Ljung (1987), it is better to calculate the sums (12.4) and (12.5) from $t = n + 1$ instead of $t = 1$. The fast algorithm of Levinson (Ljung, 1987) derives from the resulting structure of the matrix $\boldsymbol{R}$.

12.2 Gradient Methods

Gradient methods are simple and explicit. However, in the neighbourhood of the minimum, Newton-type methods such as Newton-Raphson, Gauss-Newton, quasi-Newton methods, or even Levenberg-Marquardt (Fletcher, 1991; Gill et al., 1981) are preferable with respect to their rate of convergence.

12.2.1 Gradient Method Based on a Priori Error

The deterministic model of the process is written in linear form

$$y(t) = \phi^T(t)\,\theta = \theta^T\,\phi(t) \tag{12.20}$$

where θ represents the true value of the parameter vector. The associated predictor is written in the general form

$$\hat{y}(t|\theta) = \hat{\theta}^T\,\phi(t) \tag{12.21}$$

e.g. for an ARMAX model defined by

$$y(t) = \frac{B(q)}{A(q)}\,u(t) + \frac{C(q)}{A(q)}\,e(t) \tag{12.22}$$

the observation vector of the prediction model is equal to

$$\begin{aligned}\phi(t,\theta)^T \quad = \quad & [-y(t-1),\,\ldots,\,-y(t-n_a),\,u(t-1),\,\ldots,\,u(t-n_b)\,,\\ & \epsilon(t-1,\theta),\,\ldots,\,\epsilon(t-n_c,\theta)]^T\end{aligned} \tag{12.23}$$

and the parameter vector

$$\theta^T = [a_1\,,\,\ldots\,,\,a_{n_a}\,,\,b_1\,,\,\ldots\,,\,b_{n_b}\,,\,c_1\,,\,\ldots\,,\,c_{n_c}]^T \tag{12.24}$$

Two types of predictor are distinguished (Landau, 1990):

- The a priori predictor

$$\hat{y}^o(t) = \hat{y}(t|\theta(t-1)) = \hat{\theta}^T(t-1)\,\phi(t) \tag{12.25}$$

  where we use the information contained in $\phi(t)$ until instant $t-1$, the parameters being estimated at $t-1$.
- The a posteriori predictor

$$\hat{y}(t) = \hat{y}(t|\theta(t)) = \hat{\theta}^T(t)\,\phi(t) \tag{12.26}$$

  where we still use the information contained in $\phi(t)$ until instant $t-1$, while the parameters are estimated at t.

Two types of prediction error are associated with these two predictors:

- The a priori error

$$\epsilon^o(t) = y(t) - \hat{y}^o(t) \tag{12.27}$$

- The a posteriori error

$$\epsilon(t) = y(t) - \hat{y}(t) \tag{12.28}$$

In order to avoid having to completely calculate again the parameter vector at each time step, we frequently use the structure of a parameter adaptation algorithm, through which we update the parameter vector at time t from the previous one at time $t-1$

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \Delta\theta(t) = \hat{\theta}(t-1) + f[\hat{\theta}(t-1),\,\phi(t),\,\epsilon^o(t)] \tag{12.29}$$

by minimizing an objective function J at each step with respect to the parameter vector $\hat{\theta}(t-1)$

$$\min_{\hat{\theta}(t-1)} J(t) = [\epsilon^o(t)]^2 \tag{12.30}$$

We will note that this criterion bears only on the instantaneous error.

A classical minimization method is the gradient method which includes many variants; it consists (Fig. 12.1) of getting near to the optimum by moving along the gradient in the opposite direction to the gradient, i.e. along the normal to a contour (iso-response or iso-criterion line). Recall that gradient methods are less efficient than Newton-type methods near the optimum.

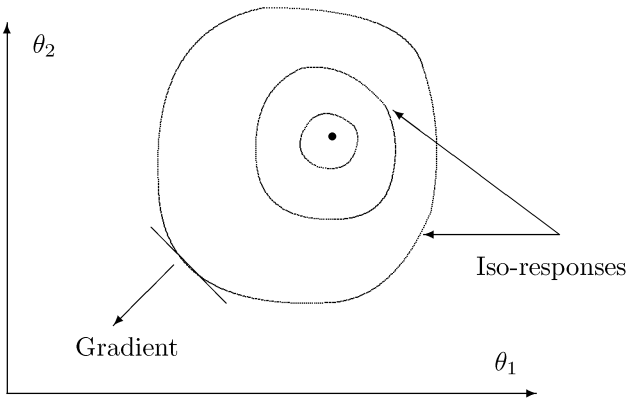


Figure 12.1: Optimization according to a gradient-type method

The gradient associated with the criterion is the vector

$$\frac{\partial J(t)}{\partial \hat{\theta}(t-1)} = 2\frac{\partial \epsilon^o(t)}{\partial \hat{\theta}(t-1)}\,\epsilon^o(t) \tag{12.31}$$

Example 12.1: **Gradient Calculation for the General Model**

We wish to calculate the gradient (Ljung, 1987) in the case of the general model (11.53) for which the following predictor is

$$\hat{y}(t|\theta) = \frac{D(q)\,B(q)}{C(q)\,F(q)}\,u(t) + \left[1 - \frac{D(q)\,A(q)}{C(q)}\right] y(t) \tag{12.32}$$

As the a priori error is equal to

$$\epsilon^o(t) = y(t) - \hat{y}^o(t) = y(t) - \hat{y}[t|\theta(t-1)] \tag{12.33}$$

to calculate the components of the gradient vector $\partial J(t)/\partial\hat{\theta}$, first calculate the partial derivatives of $\hat{y}(t|\theta)$ with respect to the successive parameters a_k, b_k, c_k, d_k, f_k of the vector θ

$$\frac{\partial}{\partial a_k}\hat{y}(t|\theta) = -\frac{D(q)}{C(q)}\,y(t-k) \tag{12.34}$$

$$\frac{\partial}{\partial b_k}\hat{y}(t|\theta) = \frac{D(q)}{C(q)\,F(q)}\,u(t-k) \tag{12.35}$$

$$\begin{aligned}\frac{\partial}{\partial c_k}\hat{y}(t|\theta) &= \frac{D(q)\,A(q)}{C(q)\,C(q)}\,y(t-k) - \frac{D(q)\,B(q)}{C(q)\,C(q)\,F(q)}\,u(t-k)\\ &= \frac{1}{C(q)}\,\epsilon^o(t-k,\theta)\end{aligned} \tag{12.36}$$

$$\begin{aligned}\frac{\partial}{\partial d_k}\hat{y}(t|\theta) &= -\frac{A(q)}{C(q)}\,y(t-k) + \frac{B(q)}{C(q)\,F(q)}\,u(t-k)\\ &= -\frac{1}{C(q)}\,v(t-k,\theta)\end{aligned} \tag{12.37}$$

$$\begin{aligned}\frac{\partial}{\partial f_k}\hat{y}(t|\theta) &= -\frac{D(q)\,B(q)}{C(q)\,F(q)\,F(q)}\,u(t-k)\\ &= -\frac{D(q)}{C(q)\,F(q)}\,w(t-k)\end{aligned} \tag{12.38}$$

The gradient vector is then calculated as

$$\frac{\partial J(t)}{\partial\hat{\theta}(t-1)} = 2\frac{\partial\epsilon^o(t)}{\partial\hat{\theta}(t-1)}\,\epsilon^o(t) = -2\frac{\partial\hat{y}(t|\theta)}{\partial\hat{\theta}(t-1)}\,(y(t)-\hat{y}(t|\theta)) \tag{12.39}$$

If we represent the linear deterministic model in the usual form

$$y(t) = \theta^T\,\phi(t) \tag{12.40}$$

the half-gradient is simply expressed as

$$\frac{1}{2}\,\frac{\partial J(t)}{\partial\hat{\theta}(t-1)} = -\phi(t)\,\epsilon^o(t) \tag{12.41}$$

so that the parameter adaptation algorithm will be written in a general manner

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P\,\phi(t)\,\epsilon^o(t) \tag{12.42}$$

where P is the adaptation gain matrix. The gain matrix must be symmetrical positive definite.

The parametric error $\tilde{\theta}(t-1)$ is the difference between the estimator $\hat{\theta}$ of the parameter vector and the true value θ of this vector

$$\tilde{\theta}(t-1) = \hat{\theta}(t-1) - \theta \tag{12.43}$$

The a priori error is then equal to

$$\epsilon^o(t) = y(t) - \hat{y}^o(t) = \theta^T\,\phi(t) - \hat{\theta}^T(t-1)\,\phi(t) = -\tilde{\theta}^T(t-1)\,\phi(t) \tag{12.44}$$

while the algorithm of parametric adaptation expresses the parametric error as

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) - P\,\phi(t)\,\phi^T(t)\tilde{\theta}(t-1) = \left[\boldsymbol{I} - P\,\phi(t)\,\phi^T(t)\right]\,\tilde{\theta}(t-1) = A(t)\,\tilde{\theta}(t-1) \tag{12.45}$$

The stability of the algorithm of parametric adaptation implies that the eigenvalues of this matrix A must be inside the unit circle.

Example 12.2: **Condition for the Gain**

Suppose that the gain matrix is a scalar positive matrix

$$P = \alpha \boldsymbol{I} \tag{12.46}$$

then, it is necessary that

$$\|\boldsymbol{I} - \alpha\, \phi(t)\, \phi^T(t)\| < 1 \tag{12.47}$$

which implies

$$\alpha < \frac{1}{\|\phi(t)\, \phi^T(t)\|} \tag{12.48}$$

12.2.2 Gradient Method Based on a Posteriori Error

Generalities

The criterion in the simple gradient method is the square of the a priori error; the improved gradient method relies on the use of the square of the a posteriori error, giving the criterion

$$\min_{\theta(t)} J(t) = [\epsilon(t)]^2 \tag{12.49}$$

where, again, the instantaneous error is concerned. The gradient is then equal to

$$\frac{\partial J(t)}{\partial \hat{\theta}(t)} = 2 \frac{\partial \epsilon(t)}{\partial \hat{\theta}(t)}\, \epsilon(t) \tag{12.50}$$

and as the a posteriori error is equal to

$$\epsilon(t) = y(t) - \hat{y}(t) = y(t) - \hat{\theta}^T(t)\, \phi(t) \tag{12.51}$$

the half-gradient becomes

$$\frac{1}{2} \frac{\partial J(t)}{\partial \hat{\theta}(t)} = -\phi(t)\, \epsilon(t) \tag{12.52}$$

and the parametric adaptation algorithm becomes

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P\, \phi(t)\, \epsilon(t) \tag{12.53}$$

Again, the gain matrix P must be symmetrical positive definite. As the a posteriori error $\epsilon(t)$ is not known, this algorithm is not physically realizable (causality principle). To realize it, we must operate a transformation depending

on the known data at $t-1$, which is realized by relating the a priori error $\epsilon^o(t)$ and the a posteriori error $\epsilon(t)$. The a posteriori error becomes equal to

$$\begin{aligned} \epsilon(t) &= y(t) - \hat{y}(t) = y(t) - \hat{\theta}^T(t)\,\phi(t) \\ &= y(t) - \hat{\theta}^T(t-1)\,\phi(t) - \left[\hat{\theta}(t) - \hat{\theta}(t-1)\right]^T \phi(t) \\ &= \epsilon^o(t) - \left[P\,\phi(t)\,\epsilon(t)\right]^T \phi(t) \end{aligned} \tag{12.54}$$

We deduce

$$\epsilon(t) = \frac{\epsilon^o(t)}{1 + \phi^T(t)\,P\,\phi(t)} \tag{12.55}$$

and the expression of the parametric adaptation algorithm in this improved version

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{P\,\phi(t)\,\epsilon^o(t)}{1 + \phi^T(t)\,P\,\phi(t)} \tag{12.56}$$

The gain matrix P must still be positive definite; on the other hand, we can verify that this algorithm is stable for any gain P. To a certain extent, the gain matrix P could be chosen in any manner by the user, provided that it is positive definite.

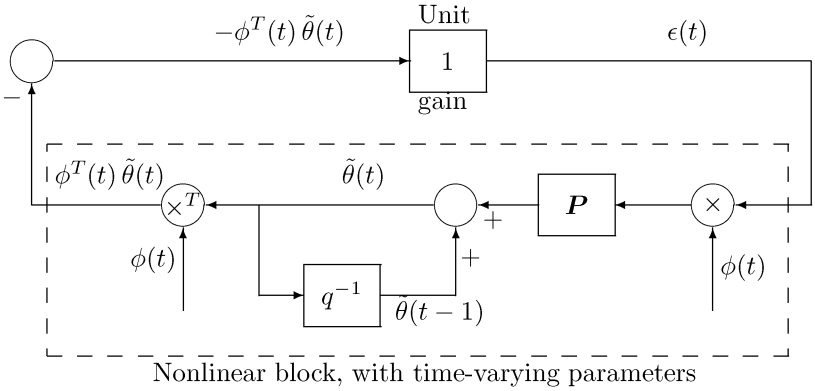


Figure 12.2: Equivalent feedback representation of the adaptive algorithm

The error of the parameter vector (θ being the true value) is defined by

$$\tilde{\theta}(t) = \hat{\theta}(t) - \theta \tag{12.57}$$

giving the a posteriori error

$$\epsilon(t) = \phi^T(t)\,\theta - \phi^T(t)\,\hat{\theta}(t) = -\phi^T(t)\,\tilde{\theta}(t) \tag{12.58}$$

and the adaptation algorithm, with respect to deviation variables,

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + P\,\phi(t)\,\epsilon(t) \tag{12.59}$$

The adaptive algorithm defined by Eqs. (12.58) and (12.59) can be represented in an equivalent manner as a feedback scheme (Fig. 12.2) (Dugard and Landau,

1990), where the signals are deviation variables. The system thus represented by the feedback scheme is not a linear system.

The stability of this algorithm does not mean that the estimated parameter vector converges towards the good value.

Extended Horizon

The fact that the a posteriori error $\epsilon^2(t)$ is minimized does not necessarily imply that the sum over an extended horizon $\sum \epsilon_i^2(i)$ is minimized; moreover, the convergence may be slow and oscillations around the minimum may occur. It is clear that the gain must be variable: large at the beginning of minimization, then small when getting closer to the minimum.

Consider the new criterion for an extended horizon based on the sum of the a posteriori errors

$$J(t) = \sum_{i=1}^{t} \left[\epsilon^2(i|\theta(t))\right]^2 = \sum_{i=1}^{t} \left[y(i) - \hat{y}(i|\theta(t))\right]^2 = \sum_{i=1}^{t} \left[y(i) - \hat{\theta}^T(t)\,\phi(i)\right]^2 \tag{12.60}$$

The solution is obtained by making the gradient equal to zero, hence

$$\frac{\partial J(t)}{\partial \hat{\theta}(t)} = -2 \sum_{i=1}^{t} \left[y(i) - \hat{\theta}^T(t)\,\phi(i)\right]\,\phi(i) = 0 \tag{12.61}$$

We deduce the estimation of the parameter vector

$$\hat{\theta}(t) = P(t) \sum_{i=1}^{t} y(i)\,\phi(i) \quad \text{with: } P^{-1}(t) = \sum_{i=1}^{t} \phi(i)\,\phi^T(i) \tag{12.62}$$

This equation thus defines the intensity of the displacement in the gradient direction corresponding to Eq. (12.53). Notice that the inverse $P^{-1}(t)$ of the matrix depends only on the measured outputs. This algorithm is non-recursive and obliges us to invert a matrix, which may be numerically delicate. It is necessary to wait for a number of steps equal to the dimension of parameter vector $\hat{\theta}$ before starting it.

12.3 Recursive Algorithms

12.3.1 Simple Recursive Least Squares

Frequently, it is desirable to have at our disposal methods that allow us to adjust on-line any model in view of the prediction, filtering or control which will be qualified as adaptive. They must display high performance with respect to stability, convergence and operation rate, as they must be performed during a sampling period. Often, they will also be used off-line.

In identification, the aim of a recursive algorithm is to find the new estimation $\hat{\theta}(t)$ from $\hat{\theta}(t-1)$ without having to perform again all the calculations.

According to the least-squares criterion, we consider again Eq. (12.62)

$$\hat{\theta}(t) = P(t) \sum_{i=1}^{t} y(i)\,\phi(i) \quad \text{with: } P^{-1}(t) = \sum_{i=1}^{t} \phi(i)\,\phi^T(i) \tag{12.63}$$

Let us write the sought estimation in a recursive form

$$\hat{\theta}(t) = P(t) \sum_{i=1}^{t} y(i)\,\phi(i) = \hat{\theta}(t-1) + \Delta\theta(t) \tag{12.64}$$

and $P^{-1}(t)$ also in a recursive form

$$\begin{aligned} P^{-1}(t) &= \sum_{i=1}^{t} \phi(i)\,\phi^T(i) \\ &= \sum_{i=1}^{t-1} \phi(i)\,\phi^T(i) + \phi(t)\,\phi^T(t) \\ &= P^{-1}(t-1) + \phi(t)\,\phi^T(t) \end{aligned} \tag{12.65}$$

Notice that $\phi(t)\,\phi^T(t) > 0$. Thus, as the inverse of the gain increases with time, the gain diminishes.

We then operate the following transformation

$$\begin{aligned} \sum_{i=1}^{t} y(i)\,\phi(i) &= P^{-1}(t)\,\hat{\theta}(t) = \sum_{i=1}^{t-1} y(i)\,\phi(i) + y(t)\,\phi(t) \\ &= P^{-1}(t-1)\,\hat{\theta}(t-1) + y(t)\,\phi(t) \\ &= P^{-1}(t-1)\,\hat{\theta}(t-1) + y(t)\,\phi(t) + \phi(t)\,\phi^T(t)\,\hat{\theta}(t-1) \\ &\quad -\phi(t)\,\phi^T(t)\,\hat{\theta}(t-1) \\ &= P^{-1}(t)\,\hat{\theta}(t-1) + \phi(t)\left[y(t) - \hat{\theta}^T(t-1)\,\phi(t)\right] \\ &= P^{-1}(t)\,\hat{\theta}(t-1) + \phi(t)\,\epsilon^o(t) \end{aligned} \tag{12.66}$$

We thus deduce the recurrence for the parameter vector

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P(t)\,\phi(t)\,\epsilon^o(t) \tag{12.67}$$

We already noticed that the gain $P(t)$ is decreasing, as we had

$$P^{-1}(t) = P^{-1}(t-1) + \phi(t)\,\phi^T(t) \tag{12.68}$$

During on-line operations, we avoid a matrix inversion because of possible numerical difficulties, and we make use of the matrix inversion lemma (11.84) with

$$A = P^{-1}(t-1)\,;\; B = \phi(t)\,;\; C = I\,;\; D = \phi^T(t) \tag{12.69}$$

It results that

$$P(t) = P(t-1) - \frac{P(t-1)\,\phi(t)\,\phi^T(t)\,P(t-1)}{1 + \phi^T(t)\,P(t-1)\,\phi(t)} \tag{12.70}$$

where it appears even more simply that the gain is decreasing. The algorithm of recursive least squares (called RLS) based on the a priori error then constitutes the set of the three following formulae

$$\left\{\begin{array}{lll} \hat{\theta}(t) = & \hat{\theta}(t-1) + P(t)\,\phi(t)\,\epsilon^o(t) \\ P(t) = & P(t-1) - \dfrac{P(t-1)\,\phi(t)\,\phi^T(t)\,P(t-1)}{1+\phi^T(t)\,P(t-1)\,\phi(t)} \\ \epsilon^o(t) = & y(t) - \hat{\theta}^T(t-1)\,\phi(t) \end{array}\right. \tag{12.71}$$

Formula (12.70) multiplied on each side of the equality sign by $\phi(t)$ gives the relation

$$P(t)\,\phi(t) = \frac{P(t-1)\,\phi(t)}{1+\phi^T(t)\,P(t-1)\,\phi(t)} \tag{12.72}$$

and relation

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{P(t-1)\,\phi(t)}{1+\phi^T(t)\,P(t-1)\,\phi(t)}\,\epsilon^o(t) = \hat{\theta}(t-1) + P(t-1)\,\phi(t)\,\epsilon(t) \tag{12.73}$$

which are useful in deducing the algorithm based on the a posteriori error.

It is better to use the a posteriori error in order to make a correction; the a posteriori error is equal to

$$\begin{array}{rcl} \epsilon(t) & = & y(t) - \hat{\theta}^T(t)\,\phi(t) \\ & = & y(t) - \hat{\theta}^T(t-1)\,\phi(t) - \left[\hat{\theta}^T(t)\,\phi(t) - \hat{\theta}^T(t-1)\,\phi(t)\right] \\ & = & \epsilon^o(t) - \left[P(t)\,\phi(t)\,\epsilon^o(t)\right]^T\phi(t) \\ & = & \epsilon^o(t) - \dfrac{\phi^T(t)\,P(t-1)\,\phi(t)\,\epsilon^o(t)}{1+\phi^T(t)\,P(t-1)\,\phi(t)} \\ & = & \dfrac{\epsilon^o(t)}{1+\phi^T(t)\,P(t-1)\,\phi(t)} \end{array} \tag{12.74}$$

The algorithm of recursive least squares (RLS) based on the a posteriori error then constitutes the set of the three following formulae

$$\left\{\begin{array}{lll} \hat{\theta}(t) = & \hat{\theta}(t-1) + P(t-1)\,\phi(t)\,\epsilon(t) \\ P(t) = & P(t-1) - \dfrac{P(t-1)\,\phi(t)\,\phi^T(t)\,P(t-1)}{1+\phi^T(t)\,P(t-1)\,\phi(t)} \\ \epsilon(t) = & \dfrac{y(t) - \hat{\theta}^T(t-1)\,\phi(t)}{1+\phi^T(t)\,P(t-1)\,\phi(t)} \end{array}\right. \tag{12.75}$$

To start the algorithm of recursive least squares at $t = 0$, a very large adaptation gain $P(0)$ must be chosen; typically, take $P(0) = G_0\boldsymbol{I}$ with $G_0 \gg 1$. This gain $P(0)$ represents a measure of the confidence of the initial estimation $\hat{\theta}(0)$.

The algorithm assumes that the system parameters are constant.

The algorithm (12.71) or (12.75) of recursive least squares provides unbiased estimations of the parameters only for ARX models in the form

$$A(q)\,y(t) = q^{-d}\,B(q)\,u(t) + e(t) \tag{12.76}$$

where the disturbance is white noise. For this model, the parameter vector θ is given by Eq. (11.9) and the observation vector ϕ by Eq. (11.14).

Interpretation According to Kalman Filter

We can consider the Kalman filter treated in Sect. 11.1.2 as a parameter estimator (Dugard and Landau, 1990; Ljung, 1987; Söderström and Stoica, 1989).

A system linear is written in state space in the form

$$\begin{array}{rcl} x(t+1) & = & A(t)\,x(t) + B(t)\,u(t) + w(t) \\ y(t) & = & C(t)\,x(t) + v(t) \end{array} \tag{12.77}$$

where v and w are uncorrelated white noises (of zero mean and respective covariance matrices R and Q).

The linear regression model is

$$\hat{y}(t|\theta) = \phi^T(t)\,\theta \tag{12.78}$$

where the parameter vector is assumed constant

$$\theta(t+1) = \theta(t) = \theta \tag{12.79}$$

so that the real output is equal to

$$y(t) = \phi^T(t)\,\theta(t) + v(t) \tag{12.80}$$

In the present case of the estimation of parameter vector θ, the state is represented by the system parameters, and we have $x(t) = \theta(t)$; $A = I$ (no dynamics for the system); $B = 0$; $w(t) = 0$ (no noise on x): $Q = 0$; $\hat{x}(t) = \hat{\theta}(t)$; $y(t) = y(t+1)$; $C = \phi^T(t)$; $v(t) = v(t+1)$; $\mathrm{E}[v(t)v^T(t)] = R$; $\hat{y}(t) = \hat{y}^o(t+1)$. The system then becomes

$$\begin{array}{rcl} \theta(t) & = & \theta(t-1) \\ y(t) & = & \phi^T(t)\,\theta(t-1) + v(t) \end{array} \tag{12.81}$$

The Kalman filter, an optimal one-step predictor that allows us to estimate the state of the system, according to Eq. (11.91), is written

$$\begin{array}{rcl} \hat{x}(t+1|t) & = & A\hat{x}(t|t-1) + A\,K(t)\,[y(t) - \hat{y}(t|t-1)] \\ \hat{y}(t|t-1) & = & C\hat{x}(t|t-1) \end{array} \tag{12.82}$$

where $K(t)$ is the Kalman gain. The estimation error of the state is equal to

$$\tilde{x}(t|t-1) = x(t) - \hat{x}(t|t-1) \tag{12.83}$$

and the estimation error covariance is

$$\mathrm{E}[\tilde{x}(t|t-1)\,\tilde{x}^T(t|t-1)] = P(t) \tag{12.84}$$

thus the gain evolution gives an idea of the variation of the estimation error covariance.

We then seek the Kalman gain $K(t)$ to minimize the estimation error covariance matrix (Eq. (11.91))

$$\begin{aligned} K^*(t) &= \arg\{\min_K \operatorname{trace}[P(t)]\} \\ &= P(t-1)\,C^T\,\left[C\,P(t-1)\,C^T + R\right]^{-1} \end{aligned} \tag{12.85}$$

By replacing $K(t)$ by its optimal value, we find the adaptation gain matrix

$$P(t) = A\,P(t-1)\,A^T + Q - A\,P(t-1)\,C^T\,\left[C\,P(t-1)\,C^T + R\right]^{-1}\,C\,P(t-1)\,A^T \tag{12.86}$$

which is a solution of a Riccati equation (see the Riccati equation (11.92) of the Kalman filter as a one-step predictor).

By applying the Kalman filter to our model ($A = I, C = \phi^T$), we thus obtain the following recursive equations based on a priori error

$$\begin{cases} \hat{\theta}(t) = & \hat{\theta}(t-1) + K(t)\,\epsilon^o(t) = \hat{\theta}(t-1) + P(t)\,\phi(t)\epsilon^o(t) \\ P(t) = & P(t-1) - \dfrac{P(t-1)\,\phi(t)\,\phi^T(t)\,P(t-1)}{R + \phi^T(t)\,P(t-1)\,\phi(t)} + Q \\ \epsilon^o(t) = & y(t) - \hat{\theta}^T(t-1)\,\phi(t) \end{cases} \tag{12.87}$$

$K(t)$ is the Kalman gain, which is equal to

$$K(t) = P(t)\,\phi(t) = \frac{P(t-1)\,\phi(t)}{R + \phi^T(t)\,P(t-1)\,\phi(t)} \tag{12.88}$$

We notice that these equations are the same as for the RLS if $R = 1$ and we assume that $Q = 0$. If Q were different from zero, the estimation error covariance matrix $P(t)$ could not tend towards 0. Given $R = 1$ and $Q = 0$, the parameter vector θ possesses a Gaussian distribution of mean $\hat{\theta}$ and covariance matrix $P(t)$. It is clear that the initialization of θ can be provided if an estimation is already known and $P(0)$ reflects the confidence that we have in this initialization.

The matrix Q has a similar role to the forgetting factor (Söderström and Stoica, 1989). When Q is "large" or the forgetting factor low, the algorithm is brisk and may follow parametric variations; when Q is "small" (close to 0), or the forgetting factor near 1, the algorithm possesses good convergence properties (is convenient for time-invariant systems). The matrix Q can be provided by the user.

Different policies of adaptation gain

The original formula

$$P^{-1}(t) = P^{-1}(t-1) + \phi(t)\,\phi^T(t) \tag{12.89}$$

can be generalized in the form

$$P^{-1}(t) = \lambda_1(t)\,P^{-1}(t-1) + \lambda_2(t)\,\phi(t)\,\phi^T(t) \tag{12.90}$$

Nevertheless, the parameters λ_1 and λ_2 must fulfil the following conditions:

- $0 < \lambda_1 < 1$, otherwise the system is unstable.
- $0 < \lambda_2 < 2$, so that the gain is decreasing.

Using the matrix inversion lemma, we obtain the adaptation gain matrix according to the following algorithm

$$P(t) = \frac{1}{\lambda_1(t)} \left[P(t-1) - \frac{P(t-1)\,\phi(t)\,\phi^T(t)\,P(t-1)}{\frac{\lambda_1(t)}{\lambda_2(t)} + \phi^T(t)\,P(t-1)\,\phi(t)} \right] \tag{12.91}$$

In order to obtain a better condition number of the matrices, it is desirable to realize a factorization (examples are Cholesky or Householder).

The recursive least-squares algorithm is then modified according to the following equations

$$\begin{cases} \hat{\theta}(t) = & \hat{\theta}(t-1) + P(t-1)\,\phi(t)\,\epsilon(t) \\ \epsilon(t) = & \dfrac{y(t) - \hat{\theta}(t-1)\,\phi(t)}{1 + \phi^T(t)\,P(t-1)\,\phi(t)} \\ P(t) = & \dfrac{1}{\lambda_1(t)} \left[P(t-1) - \dfrac{P(t-1)\,\phi(t)\,\phi^T(t)\,P(t-1)}{\frac{\lambda_1(t)}{\lambda_2(t)} + \phi^T(t)\,P(t-1)\,\phi(t)} \right] \quad \text{with: } P(0) > 0 \end{cases} \tag{12.92}$$

To start the parametric identification, two cases are possible:

- No initial information concerning the parameters is available; then, it is recommended to choose a high initial gain; the initial matrix is equal to the identity matrix times a large scalar G_o

$$P(0) = G_o\,\boldsymbol{I}\,, \quad \text{with: } G_o = 1000 \tag{12.93}$$

- Beforehand, a correct parameter estimation has been performed; the gain must be chosen small

$$P(0) = G_o\,\boldsymbol{I}\,, \quad \text{with: } G_0 \leq 1 \tag{12.94}$$

or simply equal to the estimation error covariance matrix. The matrix $P(0)$ is an image of the accuracy of the initial estimation.

Many variants bearing on the adaptation gain are possible (Dugard and Landau, 1990; Landau, 1990):

a) Decreasing gain

Set the two parameters fixed and equal to 1

$$\lambda_1 = 1\,;\ \lambda_2 = 1 \tag{12.95}$$

We get the algorithm of recursive least squares (RLS). The weight is identical for all prediction errors in criterion $J(t)$

$$J(t) = \sum_{i=1}^{t} \left[y(i) - \hat{\theta}^T(t)\phi(i) \right]^2 \tag{12.96}$$

This algorithm is convenient for identification of time-invariant systems and self-tuning controllers.

b) Fixed forgetting factor

We set

$$\lambda_2 = 1\,;\ \lambda_1(t) = \lambda_1 \tag{12.97}$$

giving the recursive relation

$$P^{-1}(t) = \lambda_1\, P^{-1}(t-1) + \phi(t)\, \phi^T(t) \tag{12.98}$$

A good choice is λ_1 close to 1 (e.g. between 0.95 and 0.99, even between 0.98 and 0.995). The criterion to be minimized is

$$J(t) = \sum_{i=1}^{t} \lambda_1^{t-i} \left[y(i) - \hat{\theta}^T(t)\phi(i) \right]^2 \tag{12.99}$$

which means that the largest weight is on the last prediction error and that it decreases from t. This gain is convenient for identification and adaptive control of slowly varying systems. The smaller the forgetting factor, the more rapidly the information contained in the observation vector is forgotten.

If the system strictly remains at the steady-state, the observation vector is zero: $\phi(t)$ =constant, and this leads to an explosion of the adaptation gain. Thus, it is necessary from time to time to excite the system or to freeze the adaptation gain.

c) Variable forgetting factor

We set

$$\lambda_2 = 1 \tag{12.100}$$

giving the recursive relation

$$P^{-1}(t) = \lambda_1(t)\, P^{-1}(t-1) + \phi(t)\, \phi^T(t) \tag{12.101}$$

We can choose

$$\lambda_1(t) = \lambda_0\, \lambda_1(t-1) + (1 - \lambda_0) \tag{12.102}$$

with $\lambda_1(0)$ and λ_0 included between 0.95 and 0.99. Thus, the convergence of $\lambda_1(t)$ towards 1 is ensured when t becomes very large (Goodwin and Sin, 1984).

The criterion to be minimized is $J(t)$

$$J(t) = \sum_{i=1}^{t} \left[\prod_{j=i}^{t} \lambda_1(j-1) \right] \left[y(i) - \hat{\theta}^T(t)\phi(i) \right]^2 \quad \text{with: } \prod_{j=1}^{t} \lambda_1(j-1) = 1 \tag{12.103}$$

The variable forgetting factor allows us to forget the beginning of the data, which may allow us to adapt to a brutal variation of the system (on one step, a small forgetting factor will be taken, then it will be set again near 1). In general, the convergence is accelerated as the adaptation gain is large during a larger horizon.

This algorithm is convenient for identification of time-invariant systems and self-tuning controllers.

d) Constant trace

The significance of a constant trace is a correction in the direction of the least squares with increase of the adaptation gain. The trace of matrix $P(t)$ is chosen as constant, giving at the initial instant

$$P(0) = \begin{bmatrix} G_o & 0 & \dots & 0 \\ 0 & G_o & 0 & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & G_o \end{bmatrix} \tag{12.104}$$

The general relation

$$P^{-1}(t) = \lambda_1(t)\, P^{-1}(t-1) + \lambda_2(t)\, \phi(t)\, \phi^T(t) \tag{12.105}$$

is used, with variable $\lambda_1(t)$ and $\lambda_2(t)$ fulfilling the conditions $0 < \lambda_1 < 1$ and $0 < \lambda_2 < 2$ and, at each step, $\lambda_1(t)$ and $\lambda_2(t)$ are adjusted so that

$$\text{trace of } P(t) = n\, G_o \tag{12.106}$$

A value of G_o included between 0.1 and 4 can be chosen.

The criterion is

$$J(t) = \sum_{i=1}^{t} \left[\prod_{j=i}^{t} \lambda_1(j-1) \right] \mu(i-1) \left[y(i) - \hat{\theta}^T(t)\phi(i) \right]^2 \tag{12.107}$$

with

$$\mu(t) = \frac{1 + \lambda_2(t)\, \phi^T(t)\, P(t-1)\, \phi(t)}{1 + \phi^T(t)\, P(t-1)\, \phi(t)} \tag{12.108}$$

This algorithm is very much used and is convenient for identification and adaptive control of variable parameter systems.

The following algorithms are combinations or simplifications of the variants which have been previously described.

e) Decreasing gain or constant trace

We go from the decreasing gain algorithm to the constant trace algorithm when the condition

$$\text{trace of } P(t) \leq n\, G_0 \tag{12.109}$$

is verified with n number of parameters and G_0 included between 0.1 and 4.

The decreasing gain or constant trace algorithm is used for identification and adaptive control of variable parameter systems in the absence of initial information.

f) Variable forgetting factor or constant trace

We go from the variable forgetting factor algorithm to the constant trace algorithm when the condition

$$\text{trace of } P(t) \leq n\, G_0 \tag{12.110}$$

is verified.

In addition to the previous discussions, the variable forgetting factor or constant trace algorithm is used for identification and adaptive control of variable parameter systems in the absence of initial information.

g/ Constant gain

Both parameters are fixed and equal to

$$\lambda_1 = 1\,;\ \lambda_2 = 0 \tag{12.111}$$

giving

$$P(t) = \ldots = P(0) = \alpha\, \boldsymbol{I} \tag{12.112}$$

In fact, we then obtain the improved gradient algorithm where the adaptation gain is scalar. Its advantage is its easy operation and its drawback a worse performance than that obtained for the forgetting factor or constant trace algorithms. It may be used for identification and adaptive control of stationary or variable parameter systems if the number of parameters of the system is low ($n \leq 3$).

h) Scalar adaptation gain

– The adaptation gain matrix is equal to the identity matrix times a variable scalar

$$P(t) = \frac{1}{c(t)}\, \boldsymbol{I} \tag{12.113}$$

– When $c(t)$ is constant, we find again the improved gradient algorithm.
– If we choose $c(t) = t$, the gain is decreasing (stochastic approximation).
– It is possible to take the recurrence

$$c(t+1) = c(t) + \phi^T(t)\, \phi(t) \tag{12.114}$$

– or still,

$$c(t+1) = \lambda_1(t)\, c(t) + \lambda_2(t) \phi^T(t)\, \phi(t) \tag{12.115}$$

with the usual conditions

$$c(0) > 0 \; ; \; 0 < \lambda_1(t) \leq 1 \; ; \; 0 \leq \lambda_2(t) < 2 \tag{12.116}$$

– If we choose $c(0) =$ trace of $P^{-1}(0)$, we obtain $c(t) =$ trace of $P^{-1}(t)$.

The performances of the scalar adaptation gain algorithm are worse than those of matrix gain algorithms and counteract their ease of use.

12.3.2 Recursive Extended Least Squares

The recursive extended least-squares (RELS) method provides unbiased parameter estimations for ARMAX models in the form

$$A(q)\, y(t) = q^{-d}\, B(q)\, u(t) + C(q)\, e(t) \tag{12.117}$$

where we thus simultaneously identify the process and the disturbance. The method gives a prediction error which must tend towards the properties of white noise. For this model, the parameter vector θ is given by Eq. (11.26) and the observation vector ϕ by Eq. (11.25). The policies of adaptation gain described by algorithm (12.92) are adapted to this method.

12.3.3 Recursive Generalized Least Squares

The method of recursive generalized least squares (RGLS) provides unbiased parameter estimations for ARARX models in the form

$$A(q)\, y(t) = q^{-d}\, B(q)\, u(t) + \frac{1}{D(q)}\, e(t) \tag{12.118}$$

In this case, we introduce an auxiliary variable

$$v(t) = A(q)\, y(t) - q^{-d}\, B(q)\, u(t) = \frac{1}{D(q)}\, e(t) \tag{12.119}$$

resulting in

$$v(t) + d_1\, v(t-1) + \ldots + d_{n_d}\, v(t-n_d) = e(t) \tag{12.120}$$

This relation corresponds to an AR model.

The predictor is introduced

$$\begin{aligned} \hat{y}(t) = \quad & -a_1\, y(t-1) - \ldots - a_{n_a}\, y(t-n_a) + b_1\, u(t-1) + \ldots + b_{n_b}\, u(t-n_b) \\ & -d_1\, v(t-1) - \ldots - d_{n_d}\, v(t-n_d) \end{aligned} \tag{12.121}$$

which will give a white prediction error

$$y(t) - \hat{y}(t) = e(t) \tag{12.122}$$

For this model, the parameter vector θ is equal to

$$\theta = [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}, d_1, \ldots, d_{n_d}]^T \tag{12.123}$$

and the observation vector ϕ is equal to

$$\phi(t,\theta)=[-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b), -v(t-1), \ldots, -v(t-n_d)]^T \tag{12.124}$$

with

$$v(t) = \hat{A}(q)\, y(t) - q^{-d}\, \hat{B}(q)\, u(t) \tag{12.125}$$

In this form, it is possible to apply the modified RLS algorithm (12.92).

Söderström and Stoica (1989) present the RGLS method applied to an ARARX model in a different form. By introducing the polynomials

$$F(q) = A(q)\, D(q)\ ; \qquad G(q) = B(q)\, D(q) \tag{12.126}$$

the process model becomes

$$F(q)\, y(t) = q^{-d}\, G(q)\, u(t) + e(t) \tag{12.127}$$

which is simply an ARX model to which the RLS, for example, can be applied, the parameter vector being equal to

$$\theta = [f_1, \ldots, f_{n_f}, g_1, \ldots, g_{n_g}]^T \tag{12.128}$$

and the observation vector ϕ being equal to

$$\phi(t,\theta) = [-y(t-1), \ldots, -y(t-n_f), u(t-1), \ldots, u(t-n_g)]^T \tag{12.129}$$

Notice that the parameter vector does not directly provide the coefficients of polynomials $A(q)$ and $B(q)$, but allows us to obtain the system transfer function $[B(q)D(q)]/[A(q)D(q)]$.

12.3.4 Recursive Maximum Likelihood

The recursive maximum likelihood (RML) method (Ljung and Söderström, 1986) can be seen (Landau, 1990) as a modification of the RELS method. The proposed presentation is that of Ljung (1987). Consider an ARMAX model

$$A(q)\, y(t) = B(q)\, u(t) + C(q)\, e(t) \tag{12.130}$$

To the observation vector $\phi(t)$ of the RELS method, it makes the vector filtered by $1/\hat{C}(q)$ correspond (it will have to be verified that the successive estimations of $C(q)$ are stable)

$$C(q)\psi(t) = \phi(t) \tag{12.131}$$

We introduce the residual

$$\bar{\epsilon}(t,\theta) = y(t) - \phi^T(t)\hat{\theta}(t) \tag{12.132}$$

As a consequence, the observation vector is equal to

$$\begin{aligned}\phi(t,\theta) = \quad & [-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b), \\ & \bar{\epsilon}(t-1,\theta), \ldots, \bar{\epsilon}(t-n_c,\theta)]^T\end{aligned} \tag{12.133}$$

and the parameter vector equal to

$$\theta = [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}, c_1, \ldots, c_{n_c}]^T \tag{12.134}$$

The predictor is equal to

$$\hat{y}(t,\theta) = \phi^T(t)\hat{\theta}(t-1) \tag{12.135}$$

and the prediction error (different from the residual)

$$\epsilon(t) = y(t) - \hat{y}(t) \tag{12.136}$$

The algorithm follows a recursive Gauss-Newton scheme, which thus necessitates us to initialize sufficiently close to the optimal value

$$\begin{aligned}R(t) \quad &= R(t-1) + \gamma(t)[\psi(t)\psi^T(t) - R(t-1)] \\ \hat{\theta}(t) \quad &= \hat{\theta}(t-1) + \gamma(t)R^{-1}(t)\psi(t)\epsilon(t)\end{aligned} \tag{12.137}$$

Frequently, in particular for RPE methods, Ljung and Söderström (1986) choose $\gamma(t) = 1/t$. It is necessary to initialize on a given horizon the RML method in order to estimate the polynomial $C(q)$, which can be done by a RELS method. Landau (1990) recommends that this horizon is at least equal to three times the number of parameters to be estimated and to make a transition by going from the RELS method to the RML method by replacing c_i by βc_i with $0 \leq \beta \leq 1$ (variable contraction factor tending towards 1). On the other hand, the stability of polynomial $C(q)$ must be verified.

12.3.5 Recursive Prediction Error Method

The recursive prediction error (RPE) method (Ljung and Söderström, 1986; Söderström and Stoica, 1989) is applied to the most general model

$$A(q)\,y(t) = \frac{B(q)}{F(q)}\,u(t) + \frac{C(q)}{D(q)}\,e(t) \tag{12.138}$$

which is written again to isolate the noise term as

$$\frac{A(q)\,D(q)}{C(q)}\,y(t) = \frac{B(q)\,D(q)}{F(q)\,C(q)}\,u(t) + e(t) \tag{12.139}$$

The best one-step predictor that can be obtained is

$$\hat{y}(t|\theta) = \left[1 - \frac{A(q)\,D(q)}{C(q)}\right]\,y(t) + \frac{B(q)\,D(q)}{F(q)\,C(q)}\,u(t) \tag{12.140}$$

The parameter vector to be estimated is equal to

$$\theta = [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}, f_1, \ldots, f_{n_f}, c_1, \ldots, c_{n_c}, d_1, \ldots, d_{n_d}]^T \quad (12.141)$$

Contrary to the RLS method, the criterion to be minimized (Ljung, 1987) in RPE

$$J(t, \theta) = \gamma(t) \frac{1}{2} \sum_{i=1}^{t} \beta_i \epsilon^2(i, \theta) \quad (12.142)$$

which is a weighted sum of the prediction errors, is not quadratic anymore for the general model with respect to θ, so that the used recursive method can be considered as an approached solving method which determines an approximation of the parameter vector $\hat{\theta}$. The algorithm will be a recursive Gauss-Newton algorithm.

Introduce the successive auxiliary variables

$$w(t, \theta) = \frac{B(q)}{F(q)} u(t) \quad (12.143)$$

$$v(t, \theta) = A(q)\, y(t) - w(t, \theta) \quad (12.144)$$

giving the prediction error

$$\epsilon(t, \theta) = y(t) - \hat{y}(t|\theta) = \frac{D(q)}{C(q)} v(t, \theta) \quad (12.145)$$

In these conditions, the observation vector is equal to

$$\begin{aligned} \phi(t, \theta) = \quad & [-y(t-1), \ldots, -y(t-n_a),\, u(t-1), \ldots, u(t-n_b), \\ & -w(t-1, \theta), \ldots, -w(t-n_f, \theta), \epsilon(t-1, \theta), \ldots, \epsilon(t-n_c, \theta), \\ & -v(t-1, \theta), \ldots, -v(t-n_d, \theta)]^T \end{aligned} \quad (12.146)$$

Also introduce the polynomial

$$G(q) = C(q)\, F(q) \quad (12.147)$$

The introduced vector ψ is the opposite of the prediction gradient, which gives the search direction

$$-\frac{d}{d\theta}\epsilon(t, \theta) = \psi^T(t, \theta) = \begin{bmatrix} \dfrac{\partial \hat{y}(t)}{\partial a_i} \\ \dfrac{\partial \hat{y}(t)}{\partial b_i} \\ \dfrac{\partial \hat{y}(t)}{\partial f_i} \\ \dfrac{\partial \hat{y}(t)}{\partial c_i} \\ \dfrac{\partial \hat{y}(t)}{\partial d_i} \end{bmatrix} = \begin{bmatrix} -\dfrac{D(q)}{C(q)} y(t-i) \\ \dfrac{D(q)}{G(q)} u(t-i) \\ -\dfrac{D(q)}{G(q)} w(t-i) \\ \dfrac{1}{C(q)} \epsilon(t-i) \\ -\dfrac{1}{C(q)} v(t-i) \end{bmatrix} \quad (12.148)$$

The algorithm of RPE method is then described by the set of following equations (Ljung and Söderström, 1986)

$$
\begin{array}{lcl}
\epsilon(t) & = & y(t) - \hat{y}(t) \\
R(t) & = & R(t-1) + \gamma(t)[\psi(t)\psi^T(t) - R(t-1)] \\
\hat{\theta}(t) & = & \hat{\theta}(t-1) + \gamma(t)R^{-1}(t)\psi(t)\epsilon(t) \\
w(t) & = & \hat{b}_1 u(t-1) + \ldots + \hat{b}_{n_b} u(t-n_b) - \hat{f}_1 w(t-1) - \ldots - \hat{f}_{n_f} w(t-n_f) \\
v(t) & = & y(t) + \hat{a}_1 y(t-1) + \ldots + \hat{a}_{n_a} y(t-n_a) - w(t) \\
\bar{\epsilon}(t) & = & v(t) + \hat{d}_1 v(t-1) + \ldots + \hat{d}_{n_d} v(t-n_d) \\
 & & -\hat{c}_1 \bar{\epsilon}(t-1) - \ldots - \hat{c}_{n_c} \bar{\epsilon}(t-n_c) \\
\phi(t+1) = & & [-y(t), \ldots, -y(t-n_a+1), u(t), \ldots, u(t-n_b+1), \\
 & & -w(t), \ldots, -w(t-n_f+1), \bar{\epsilon}(t), \ldots, \bar{\epsilon}(t-n_c+1), \\
 & & -v(t), \ldots, -v(t-n_d+1)]^T \\
\hat{y}(t+1) = & & \hat{\theta}(t)\phi(t+1) \\
\tilde{y}(t) & = & y(t) + \hat{d}_1 y(t-1) + \ldots + \hat{d}_{n_d} y(t-n_d) \\
 & & -\hat{c}_1 \tilde{y}(t-1) - \ldots - \hat{c}_{n_c} \tilde{y}(t-n_c) \\
\tilde{u}(t) & = & u(t) + \hat{d}_1 u(t-1) + \ldots + \hat{d}_{n_d} u(t-n_d) \\
 & & -\hat{g}_1 \tilde{u}(t-1) - \ldots - \hat{g}_{n_g} \tilde{u}(t-n_g) \\
\tilde{w}(t) & = & w(t) + \hat{d}_1 w(t-1) + \ldots + \hat{d}_{n_d} w(t-n_d) \\
 & & -\hat{g}_1 \tilde{w}(t-1) - \ldots - \hat{g}_{n_g} \tilde{w}(t-n_g) \\
\tilde{\epsilon}(t) & = & \bar{\epsilon}(t) - \hat{c}_1 \tilde{\epsilon}(t-1) - \ldots - \hat{c}_{n_c} \tilde{\epsilon}(t-n_c) \\
\tilde{v}(t) & = & v(t) - \hat{c}_1 \tilde{v}(t-1) - \ldots - \hat{c}_{n_c} \tilde{v}(t-n_c) \\
\psi(t+1) = & & [-\tilde{y}(t), \ldots, -\tilde{y}(t-n_a+1), \tilde{u}(t), \ldots, \tilde{u}(t-n_b+1), \\
 & & -\tilde{w}(t), \ldots, -\tilde{w}(t-n_f+1), \tilde{\epsilon}(t), \ldots, \tilde{\epsilon}(t-n_c+1), \\
 & & -\tilde{v}(t), \ldots, -\tilde{v}(t-n_d+1)]^T
\end{array}
\tag{12.149}
$$

Notice a difference between this algorithm and the immediately previous equations: the prediction error $\epsilon(t)$ (a priori) is replaced by the residual $\bar{\epsilon}(t)$ (a posteriori).

By using $P(t) = \gamma(t)R^{-1}(t)$, a factor of $\psi(t)\epsilon(t)$ in the recursive formula of $\theta(t)$, we notice that it is useful to realize a decomposition of this matrix to ensure that it remains positive definite along the iterations. Ljung (1987) proposes the Bierman algorithm to replace P by a product UDU^T (U upper triangular matrix and D diagonal). The matrix P is initialized as $P = P_0 \boldsymbol{I}$ with large P_0. The forgetting factor $\gamma(t)$ can be chosen, as in RML, to be equal to $1/t$.

Along the iterations, a stability test concerning polynomials C and F must also be performed.

As the RPE method is very general, it is interesting to study some particular cases:

- If $F(q) = C(q) = D(q) = 1$ (ARX model), then we obtain the recursive least squares (RLS).
- If $F(q) = D(q) = 1$ (ARMAX model), then we obtain the recursive maximum likelihood (RML).

- If $A(q) = C(q) = D(q) = 1$ corresponding to the following model

$$y(t) = \frac{B(q)}{F(q)}\, u(t) + v(t) \tag{12.150}$$

we obtain an output error method (Dugard and Landau, 1980).

In general, the recursive prediction error (RPE) method converges faster than the recursive extended least squares (RELS) (Söderström and Stoica, 1989).

It must be noted that the extended Kalman filter, which is applicable to nonlinear models, can be used to estimate the parameters of the general model (Ljung, 1987). However, the RPE method ensures the convergence contrary to the extended Kalman filter.

12.3.6 Instrumental Variable Method

In the previous chapter concerning the principle of instrumental variable method, two important properties were stated:

- The estimator is given by

$$\theta^{VI} = \left[\frac{1}{N}\sum_{t=1}^{N} \zeta(t)\,\phi^T(t)\right]^{-1} \frac{1}{N}\sum_{t=1}^{N} \zeta(t)\, y(t) \tag{12.151}$$

provided that the inverse matrix exists.

- The instrumental variable vector must verify that

$$\begin{cases} \mathrm{E}\left[\zeta(t)\,\phi^T(t)\right] & \text{is nonsingular} \\ \mathrm{E}\left[\zeta(t)\, v_0(t)\right] = 0 \end{cases} \tag{12.152}$$

thus, contrary to the methods based on the minimization of the sum of the squares of the prediction errors, the instrumental variable method is based on the uncorrelation of the instrumental variables and the regression variables.

The formula (12.151) suggests that it is possible to apply an algorithm that is similar to the recursive least-squares algorithm, so that the algorithm of the instrumental variable method (Ljung and Söderström, 1986) will be the following

$$\begin{cases} \hat{\theta}(t) = & \hat{\theta}(t-1) + L(t)\left[y(t) - \hat{\theta}^T(t-1)\,\phi(t)\right] \\ L(t) = & P(t)\zeta(t) = \dfrac{P(t-1)\zeta(t)}{1 + \phi^T(t)\,P(t-1)\,\zeta(t)} \\ P(t) = & P(t-1) - \dfrac{P(t-1)\zeta(t)\,\phi^T(t)\,P(t-1)}{1 + \phi^T(t)\,P(t-1)\,\zeta(t)} \end{cases} \tag{12.153}$$

This method presents many variants related to the choice of the instrumental variables ζ, of which a frequent possibility is cited in Sect. 11.4.5. Landau (1990) recommends initializing the method by the RLS method over an horizon equal to at least three times the number of parameters to be estimated.

12.3.7 Output Error Method

This method, in the same way as the instrumental variable method, is based on the uncorrelation of the observations and the prediction errors (Landau, 1990). The vectors θ and $\phi(t)$ are given by Eqs. (11.48) and (11.49). An algorithm similar to RLS can be used. As the vector $\phi(t)$ takes into account the model outputs instead of the actual outputs, this method is less sensitive to the output noise than the RLS method.

12.4 Algorithm Robustification

It is not sufficient to have chosen a good parametric adaptation algorithm. Some recommendations must be used in view of an efficient on-line use (Dugard and Landau, 1990; Isermann, 1991; Landau, 1988, 1990; Middleton and Goodwin, 1990).

In the case of RLS, to ensure parameter's convergence towards their true value, the regression vector $\phi(t)$ must fulfil the following condition (Middleton and Goodwin, 1990) of sufficient variation

$$\text{There exists } t_1 \text{ and } \epsilon \text{ such that } \quad \sum_{i=t}^{t_1} \phi(i)\phi^T(i) \geq \epsilon \boldsymbol{I} \qquad \forall t \tag{12.154}$$

and that the excitation is persistent: the input u must be such that the output varies sufficiently; the input is persistently exciting. These conditions are, in fact, desired in general.

- The input and output signals having to possess a zero mean, their steady-state component must be eliminated, which is realized by subtracting their time mean.
- An anti-aliasing filter must be incorporated in the system (refer to Shannon theorem in signal processing) in order to avoid distortion.
- The signals in the significant bandwidth must be favoured:
  - Filter the low-frequency signals to avoid the drifts and load disturbances, and then work on the filtered signals.
  - Filter the high-frequency signals to eliminate the measurement noise and the unmodelled dynamics.
- Data filtering:
  Suppose that the model before filtering includes a noise as

$$y(t) = \theta^T \phi(t) + \eta(t) \tag{12.155}$$

  - Case of coloured noise:
    The coloured noise η is defined from its spectral characteristics, as filtered white noise e

$$\eta(t) = \frac{C(q)}{D(q)} e(t) \tag{12.156}$$

We prefilter the data to form

$$\phi_f = \frac{D(q)}{C(q)}\,\phi\;; \qquad y_f = \frac{D(q)}{C(q)}\,y \tag{12.157}$$

so that the filtered estimation model simply makes white noise appear

$$y_f(t) = \theta^T\,\phi_f(t) + e(t) \tag{12.158}$$

Then it will suffice to apply a recursive algorithm to the filtered variables y_f and ϕ_f.

– Case of a disturbance:
In a process, it is necessary first to make an analysis of the most frequent disturbances. Then, the disturbance $\eta(t)$ can be modelled using a polynomial $S(q)$ depending on the disturbance type (Middleton and Goodwin, 1990) such as

$$S(q)\eta(t) = 0 \tag{12.159}$$

A stable polynomial $Q(q)$ is chosen, with degree larger or equal to that of S, and the data are filtered

$$\phi_f = \frac{S(q)}{Q(q)}\,\phi\;; \qquad y_f = \frac{S(q)}{Q(q)}\,y \tag{12.160}$$

giving the filtered model without disturbance

$$y_f(t) = \theta^T\,\phi_f(t) \tag{12.161}$$

The polynomial $S(q)$ is the internal model of the disturbance; e.g. for a step disturbance, which should always be taken into account

$$S(q) = 1 - q^{-1} \tag{12.162}$$

and for a sinusoidal disturbance

$$S(q) = 1 - 2\cos(\omega_0 T_s)q^{-1} + q^{-2} \tag{12.163}$$

- Data normalization:
Consider the following process model

$$A(q)\,y(t) = q^{-d}\,B(q)\,u(t) + w(t) \tag{12.164}$$

$w(t)$ represents the unmodelled response of the actual system. We assume that this unmodelled part is low in mean with respect to the modelled part.

To ensure the bounding of $w(t)$, the data are normalized with the following normalization factor

$$\eta(t) = \mu\,\eta(t-1) + g\,\max[\phi^T\,\phi(t), \eta_o] \text{ with: } 0 \le \mu < 1\;;\; g > 0\;;\; \eta_0 > 0 \tag{12.165}$$

giving the normalized variables

$$\bar{u}(t) = \frac{u(t)}{\sqrt{\eta(t)}} \; ; \; \bar{y}(t) = \frac{y(t)}{\sqrt{\eta(t)}} \; ; \; \bar{\phi}(t) = \frac{\phi(t)}{\sqrt{\eta(t)}} \tag{12.166}$$

and the new model

$$A(q)\,\bar{y}(t) = q^{-d}\,B(q)\,\bar{u}(t) + \bar{w}(t) \tag{12.167}$$

hence

$$\bar{y}(t) = \theta^T\,\bar{\phi}(t) + \bar{w}(t) \tag{12.168}$$

The adaptation algorithm with the normalized data will be the same as the one with the raw data. We obtain

$$\begin{aligned} \hat{\theta}(t) &= \hat{\theta}(t-1) + \frac{P(t-1)\,\bar{\phi}(t)\,\epsilon^o(t)}{1+\bar{\phi}^T(t)\,P(t-1)\,\bar{\phi}(t)} \\ &= \hat{\theta}(t-1) + \frac{P(t-1)\,\phi(t)\,\epsilon^o(t)}{\eta(t)+\phi^T(t)\,P(t-1)\,\phi(t)} \end{aligned} \tag{12.169}$$

$$P(t) = \frac{1}{\lambda_1(t)}\left[P(t-1) - \frac{P(t-1)\,\phi(t)\,\phi^T(t)\,P(t-1)}{\frac{\lambda_1(t)}{\lambda_2(t)}\,\eta(t)+\phi^T(t)\,P(t-1)\,\phi(t)}\right] \tag{12.170}$$

- Case of time-varying parameter systems:
  In the case of systems with time-varying parameters $\theta(t)$, there exists no stationary value θ. This technique can also be used to take into account unmodelled dynamics.

  We use the following prediction of the parameter vector

$$\hat{\theta}(t) = \sigma\,\hat{\theta}(t-1) + P(t-1)\,\phi(t)\,\epsilon(t) \text{ with : } 0.95 < \sigma < 0.99 \tag{12.171}$$

  which amounts to replacing the integrator bearing on $\theta(t)$ by a first-order filter; σ is called a contraction factor.

  Let $\tilde{\theta}(t)$ be the error committed on the parameter vector (of unknown theoretical value θ)

$$\tilde{\theta}(t) = \theta(t) - \theta \tag{12.172}$$

  With the contraction factor, we obtain the error

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + P(t-1)\,\phi(t)\,\epsilon(t) - (1-\sigma)\,\theta \tag{12.173}$$

  The term $(1-\sigma)\,\theta$ constitutes a drift error, which is the counterpart due to the better stability.
- Various techniques:
  Different other techniques exist that allow us to improve the behaviour of parametric adaptation algorithms, e.g.
  - The use of the "dead zone": consisting of freezing the value of the parameter vector when the prediction error is larger than the noise amplitude.

– The use of a projection sphere: the parameters are forced to remain inside a domain by projection on the boundaries if their estimation is situated outside.
– Factorization of the adaptation gain: the objective is to ensure that the gain $P(t)$ is a positive definite matrix for all t (Cholesky decomposition, etc.).

12.5 Validation

In the case of off-line identification, the set of input-output data must be separated into two parts: the first one being used for identification, the second for validation.

The aim of validation is to judge the quality of the identification method and the quality of the obtained model. The model must present a compromise between its fitting to input-output data and its complexity, which must not be too important (risk of over-parameterization).

Two classes of identification methods have been introduced:

- Methods based on the hypothesis that the prediction error must have the properties of white noise. This is the case of: Recursive Least Squares (RLS), Recursive Extended Least Squares (RELS), Recursive Generalized Least Squares (RGLS), Recursive Maximum Likelihood (RML), Recursive Prediction Error Method (RPE).

  To verify if the prediction error approximates white noise, the autocorrelation of the prediction error is estimated

$$R(i) \approx \frac{1}{N} \sum_{k=1}^{N} \epsilon(k)\epsilon(k-i) \tag{12.174}$$

  and normalized

$$Rn(i) = \frac{R(i)}{R(0)} \tag{12.175}$$

  Theoretically, if it were effectively white noise, we would obtain $Rn(0) = 1$ and $|Rn(i)| = 0, i > 0$. In fact, considering that the autocorrelation follows a Gaussian distribution of zero mean and standard deviation $1/\sqrt{N}$, a statistical criterion at 5% α-level is that the autocorrelation must then be such that

$$Rn(0) = 1 \; ; \qquad |Rn(i)| \leq \frac{1.96}{\sqrt{N}} \quad , i > 0 \tag{12.176}$$

  with: N number of measurements.
- The methods that are based on the hypothesis that the observation vector and the prediction error are not correlated. This is the case of instrumental variable and output error methods (Landau, 1990; Ljung, 1987; Ljung and Söderström, 1986; Söderström and Stoica, 1989).

12.6 Input Sequences for Identification

To obtain the convergence of the parameter vector of the model towards the true value, which is to realize good identification, it is important to choose an input sufficiently rich in frequencies. A type of input frequently used for its richness is the pseudo random binary sequence (PRBS) (Landau, 1990), which is formed by rectangular impulses of variable duration and of amplitude that is alternatively $+1$ or -1, the mean amplitude for the full sequence is practically zero, which allows us to limit the amplitude of the disturbance introduced on the process during identification. Other types of inputs can be used, such as rectangular impulses of variable length and amplitude, allowing us to easily detect non-linearities. A PRBS-type input approximates white noise (recall that white noise is characterized by a constant spectral density). In the case of the PRBS, the spectral density will be practically constant until about 0.3 times the sampling frequency f_e.

A program generating a PRBS can be realized according to the scheme in Fig. 12.3.

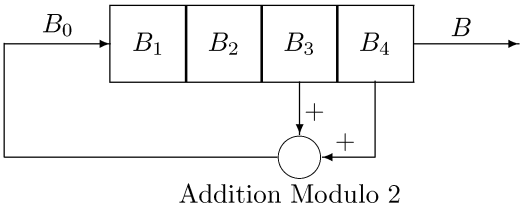


Figure 12.3: Generation of a pseudo random binary sequence for $N = 4$

If an integer N is chosen, the maximum length of the PRBS is $2^N - 1$ (then the sequence is periodical).

In Fig. 12.3, each box or register B_i represents a bit (value 0 or 1). The PRBS is represented by the values of the output B equal to B_4 along the iterations. The addition modulo 2 of B_4 and B_3 gives B_0 equal to B_1 (if B_4 and B_3 are both equal to 0 or 1, B_0 is 0; if B_4 and B_3 have different values, B_0 is 1). To go from an iteration to the following iteration, it suffices to translate the values in the registers: B_1 takes the value of B_0 which has just been calculated; B_2 takes the value of B_1, ..., B_4, thus B takes the value of B_4. In the case of Fig. 12.3, $N = 4$ was chosen, which corresponds to a maximum length of the sequence equal to 15.

Table 12.1 gives the characteristic polynomials of the PRBS, which allows us to build schemes such as Fig. 12.3 for different lengths of PRBS. In the case where $N = 4$, the addition modulo 2 thus gives

$$[R^4 \oplus R^3 \oplus R^0]B_1 = 0 \tag{12.177}$$

where R is a backward shift operator such that

$$R^j B_i = B_{i+j} \tag{12.178}$$

Table 12.1: Characteristic polynomials of PRBS of maximum length

N	Characteristic polynomial	Maximum length	Output periodic sequence
2	$R^2 \oplus R^1 \oplus R^0$	3	110
3	$R^3 \oplus R^2 \oplus R^0$	7	1110010
4	$R^4 \oplus R^3 \oplus R^0$	15	111100010011010
5	$R^5 \oplus R^3 \oplus R^0$	31	11111000110111010100001001011 00
6	$R^6 \oplus R^5 \oplus R^0$	63	11111100000100001100010100...

The addition rule (12.177) can be written in an equivalent manner

$$R^0\, B_1 = R^4\, B_1 \oplus R^3\, B_1 \tag{12.179}$$

thus

$$R^1\, B_0 = B_5 \oplus B_4 = R^1\, [B_4 \oplus B_3] \tag{12.180}$$

hence

$$B_0 = B_4 \oplus B_3 \tag{12.181}$$

which indeed corresponds to the concerned scheme in Fig. 12.3.

If we begin by the sequence 1111 (each register contains the value 1), the output of the obtained sequence of maximum length with $N = 4$ will be

$$1111\ 000\ 1\ 00\ 11\ 0\ 1\ 0 \tag{12.182}$$

and the PRBS would be (replacing values 0 by -1)

$$1\,1\,1\,1 \quad -1\,-1\,-1\ 1 \quad -1\,-1\ 1\,1 \quad -1\ 1 \quad -1 \tag{12.183}$$

Let T_s be the sampling period. To allow correct identification, at least one of the rectangular pulses must have a length larger than the rise time τ characteristic of the process. As the maximum length of a rectangular pulse of the PRBS is N sampling periods, the following rule results

$$N\, T_s > \tau \tag{12.184}$$

On the other hand, the duration of the identification t_{id} must, of course, be larger than the duration of the PRBS to cover all the frequency spectrum generated by the PRBS, thus a second rule results

$$t_{id} > (2^N - 1)\, T_s \tag{12.185}$$

As the first rule could end at too high values of N leading to an excessive duration of the PRBS, often an under-multiple of the sampling frequency is used as the basis frequency for the PRBS

$$\nu_{PRBS} = \frac{\nu_e}{p} \quad \text{with: } p = 1, 2, \ldots \tag{12.186}$$

giving the new rule, replacing the first rule

$$p\,N\,T_s > \tau \tag{12.187}$$

By using $p > 1$, we reduce the frequency spectrum of the PRBS where the spectral density is approximately constant. In general, this is not too important, provided the process has a low bandwidth.

12.7 Identification Examples

12.7.1 Academic Example of a Second-Order System

Example 12.3: **Identification of a Second-Order System**

To display some identification characteristics, first consider a continuous system, of transfer function

$$G(s) = \frac{5}{9s^2 + 3s + 1} \tag{12.188}$$

which is underdamped second-order. This transfer function has been discretized by a zero-order holder with a sampling period $T_s = 0.5$, giving the discrete transfer function

$$\frac{B(q)}{A(q)} = \frac{0.06559q + 0.06204}{q^2 - 1.8210q + 0.8465} \tag{12.189}$$

The response to a unit step is represented in Fig. 12.4. Now, we will only consider this discrete system.

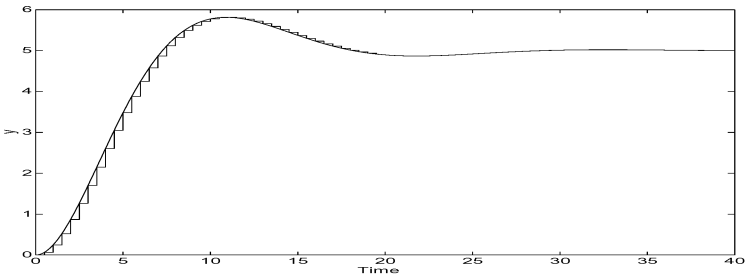


Figure 12.4: Response of a continuous second-order system and the discretized system to a unit step input

An ARMAX model corresponding to this transfer function $B(q)/A(q)$ was built by adding a noise term $C(q)$

$$\begin{aligned} y(t) - 1.8210y(t-1) + 0.8465y(t-2) = \quad & 0.06559u(t-1) + 0.06204u(t-2) \\ +e(t) - 1.1e(t-1) + 0.3e(t-2) & \end{aligned} \tag{12.190}$$

where $e(t)$ is Gaussian white noise of fixed standard deviation σ_e (here, $\sigma_e = 1$).

This model was subjected to a pseudo random binary sequence (PRBS) of characteristic number $N = 5$, whose base period was taken to be equal to four times the sampling period. This model will be subsequently called simulated or pseudo-experimental system. The input and the output of the simulated system are represented with respect to normalized time (one unit = one sampling period) in Fig. 12.5.

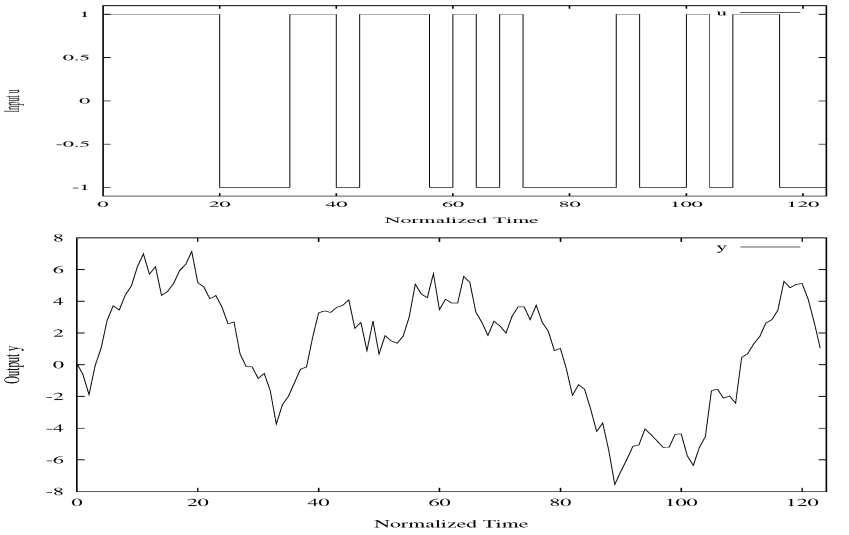


Figure 12.5: Pseudo random binary sequence (top) and response (bottom) of the discrete system to this PRBS

The input-output data thus obtained were first identified by a recursive extended least-squares method with an ARMAX model with the same number of parameters by using a constant forgetting factor $\lambda = 0.99$.

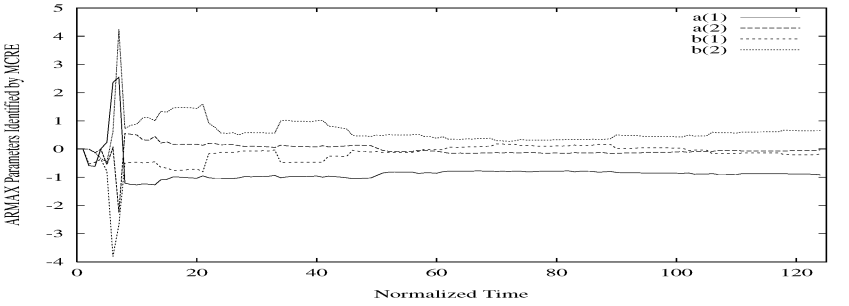


Figure 12.6: Evolution of the parameters of polynomials $A(q)$ and $B(q)$ of the identified ARMAX model (RELS method) in the case where $\sigma_e = 1$

After a transient period, the parameters converge rather quickly (Fig. 12.6). The output obtained from the identified model can be compared to the simulated experimental output in Fig. 12.7. In spite of the important noise which concerned the simulated system, the dynamics of the identified model is near that of the simulated system. The prediction error (Fig. 12.8) was calculated and we must verify that it presents properties close to white noise. The autocorrelation of the prediction error (Fig. 12.9) is included in the confidence interval at 5% α-level.

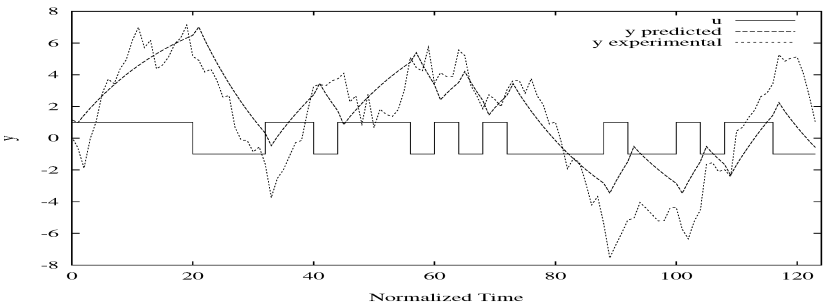


Figure 12.7: Comparison of the predicted output and the experimental output (RELS) in the case of a standard deviation of the simulation model $\sigma_e = 1$

Several criteria have been calculated to characterize the identification performance:

- The variance of the prediction error $\mathcal{C}_1 = \sigma_\epsilon^2$.
- Akaike's criterion of theoretical information (Söderström and Stoica, 1989):

$$\mathcal{C}_2 = N \log(\sigma^2) + 2d_m \tag{12.191}$$

- Akaike's criterion of final prediction error (Ljung, 1987) equal to

$$\mathcal{C}_3 = \frac{1 + d_m/N}{1 - d_m/N}\sigma_\epsilon^2 \tag{12.192}$$

  where d_m is the model dimension, thus the number of parameters to be determined.

The evolution of these criteria was studied with respect to:

- The standard deviation σ_e of the noise $e(t)$ used to generate the simulated system outputs. Table 12.2 clearly shows that, before realizing identification, it is useful to filter the data or even better still, to realize direct filtering of the sensors of the system.
  The values of the final parameters (the length of the input sequence is only equal to 124) obtained for the model identified by the RELS method are given for these different values of σ_e, as well as for the RML and RPE methods. Of course, the smaller the noise, the closer the final parameters of $A(q)$ and $B(q)$ are to their theoretical value. Even when the final parameters

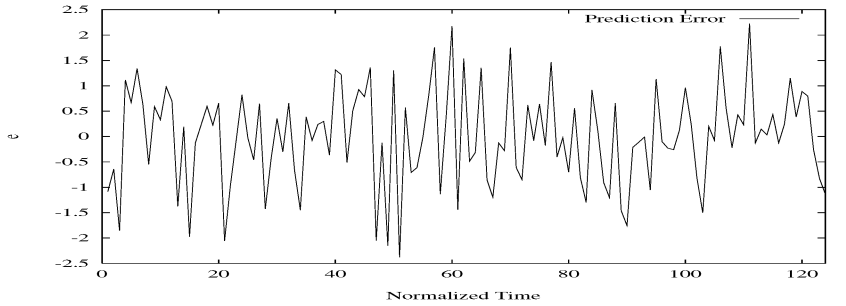


Figure 12.8: Prediction error (RELS) in the case where $\sigma_e = 1$

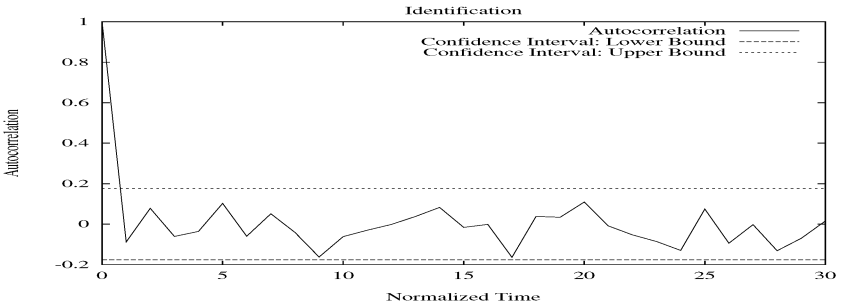


Figure 12.9: Autocorrelation of the prediction error (RELS) in the case where $\sigma_e = 1$

Table 12.2: Influence of the standard deviation of the error of the ARMAX model on the identification result expressed by criteria $\mathcal{C}_i$

RELS method (with $\lambda = 0.99$)	$\sigma_e = 1$	$\sigma_e = 0.5$	$\sigma_e = 0.25$	$\sigma_e = 0.10$
$\mathcal{C}_1$	0.912	0.269	0.0830	0.0138
$\mathcal{C}_2$	8.531	−143	−289	−511.2
$\mathcal{C}_3$	1.072	0.317	0.0975	0.0162

are far from their theoretical value, the dynamics of the identified model is correct. On the other hand, of course, the predicted output is much closer to the experimental output (Fig. 12.11) when the noise is smaller.

– The value of the constant forgetting factor λ (Table 12.4): the increase of λ beyond 0.99 brings little improvement, on the other hand its decrease leads to a neat deterioration.

On the other hand, different versions of the RPE method taking into account different forgetting factors or different decompositions of the matrix P, have given results rather close to previous RPE method. Figure 12.10 shows the predicted output by means of the model identified by the RPE method in the case where $\sigma_e = 1$.

To improve the estimation results, it is possible to proceed in several runs: after

Table 12.3: Values of the parameters of the simulated model and identified models. Value of the prediction error criterion

	a_1	a_2	b_1	b_2	Criterion $\mathcal{C}_1$
Theoretical parameters	−1.8210	0.8465	0.06559	0.06020	
RELS ($\sigma_e = 1$)	−0.91552	−0.02954	−0.18114	0.65439	0.912
RML ($\sigma_e = 1$)	−0.89830	−0.04031	−0.18298	0.66012	0.913
RPE ($\sigma_e = 1$)	−1.05677	0.10273	−0.29681	0.80870	0.985
RELS ($\sigma_e = 0.5$)	−1.18897	0.23026	−0.04587	0.37495	0.269
RELS ($\sigma_e = 0.25$)	−1.50796	0.54010	0.02823	0.19856	0.0830
RELS ($\sigma_e = 0.10$)	−1.77681	0.80251	0.05939	0.08889	0.0138
RELS ($\sigma_e = 1$) with 20 runs	−1.38486	0.42463	−0.12837	0.47795	0.8564

Table 12.4: Influence of the forgetting factor in the RELS method on the identification result

RELS method (with $\sigma_e = 1$)	$\lambda = 0.95$	$\lambda = 0.99$	$\lambda = 0.995$
$\mathcal{C}_1$	1.014	0.912	0.910
$\mathcal{C}_2$	21.68	8.531	8.279
$\mathcal{C}_3$	1.191	1.072	1.069

having used the data set for a first parametric estimation, the last estimations are kept and used as initialization for a new estimation run. In this manner, the recursive identification approaches off-line identification (Ljung and Söderström, 1986). This method was applied to a set of initial small size ($N = 124$) which had already been used and resulted in an improvement of the criterion $\mathcal{C}_1$ from 0.9117 to 0.856 and the parameters (Table 12.3) in a significant manner (Fig. 12.12).

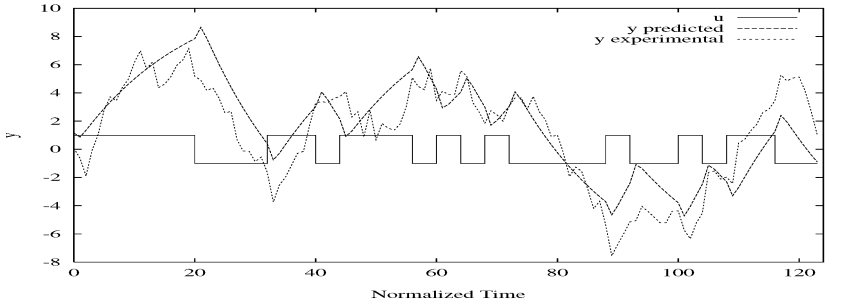


Figure 12.10: Comparison of the predicted output and of the experimental output (RPE method) in the case where $\sigma_e = 1$

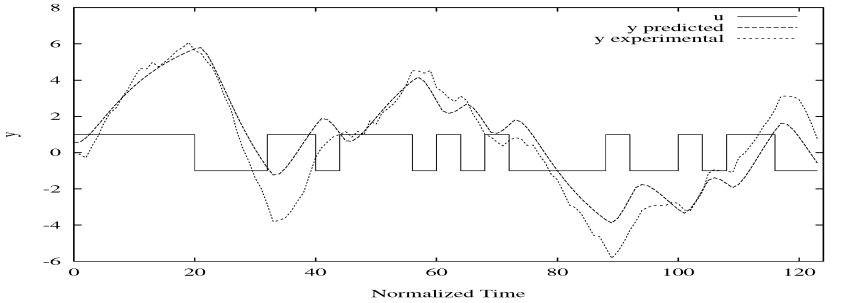


Figure 12.11: Comparison of the predicted output and the experimental output (RELS method) in the case where $\sigma_e = 0.25$

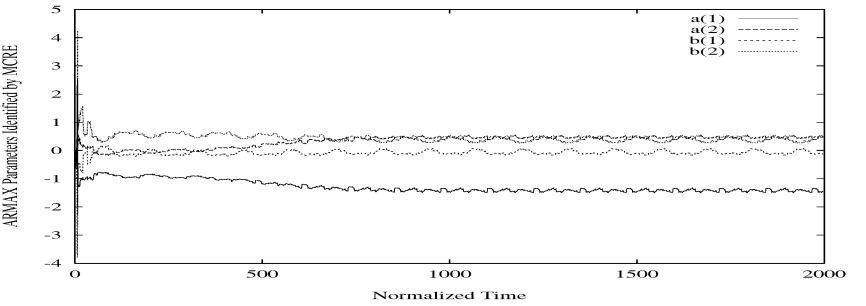
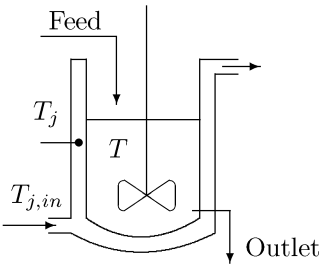


Figure 12.12: Evolution by several runs of the parameters of polynomials $A(q)$ and $B(q)$ of the identified ARMAX model (RELS method) in the case where $\sigma_e = 1$

12.7.2 Identification of a Simulated Chemical Reactor

Example 12.4: Identification of a Chemical Reactor



A chemical reactor with a jacket was simulated by means of a nonlinear knowledge model (Sect. 19.2). The chemical reaction is considered as a disturbance and only the thermal behaviour of the reactor is analyzed.

Concerning control, the input is the position of a three-way valve which drives the temperature of the heating-cooling fluid entering in the jacket, and the output is the temperature of the reactor contents. This corresponds to a lab pilot reactor. A pseudo-random binary sequence is imposed on the input after the steady-state has been reached for an input equal to the mean amplitude of the PRBS. The base period of the PRBS is 250 s and the sampling period is 5 s. The first model identified by the RELS method with a forgetting factor equal to 0.99 is an ARMAX model with degrees of $A(q), B(q), C(q)$ all equal to 1, thus only one coefficient was identified for each polynomial. We notice that the thermal dynamics of the reactor is very correctly respected (Fig. 12.13). If the degree of all polynomials is increased to 2, the difference becomes very small and non-perceptible in the figure. The reactor has indeed a behaviour close to a first-order system. An ARX model identified by RLS also gives nearly identical results.

The same knowledge model for the reactor was used, but the output temperature was polluted by realistic Gaussian white noise of standard deviation 0.5. The identification was performed in a parallel manner. In this case, with an identical number of parameters n_a and n_b, the ARX model gives slightly less satisfactory results than the ARMAX model. On the other hand, the ARMAX model (denoted by ARMAX 2-2) with the degrees of $A(q), B(q), C(q)$ all equal to 2 is very slightly worse than the ARMAX model (denoted by ARMAX 1-1) with the degrees of $A(q), B(q), C(q)$ equal to 1 (Fig. 12.14). This is confirmed by the values of the criteria obtained for the reactor with noise, gathered in Table 12.5. All models, in the cases without and with noise, have in common possession of a pole $z \approx 0.976$.

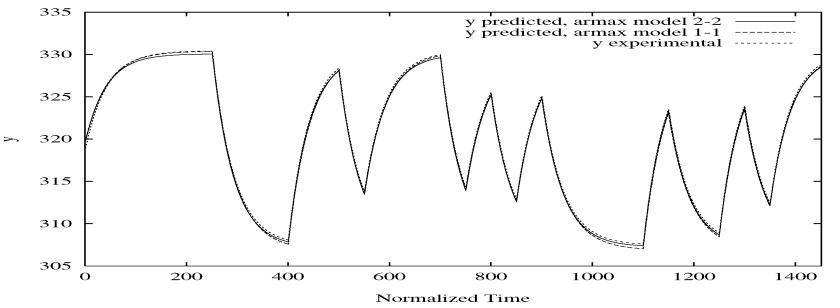


Figure 12.13: Comparison of the temperature predicted according to an ARMAX model identified by the RELS method and of the experimental temperature of the reactor contents (case without noise)

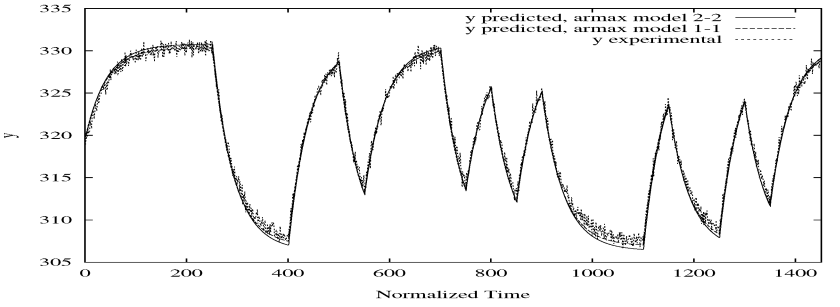


Figure 12.14: Comparison of the temperature predicted from an ARMAX model identified by the RELS method and of the experimental temperature of the reactor contents (case of the output with noise). The line of the ARMAX 1-1 model is slightly closer to the experimental values than the line of the ARMAX 2-2 model

Table 12.5: Values of the identification criteria for a reactor with output noise

Reactor with noise	ARX 1-1	ARX 2-2	ARMAX 1-1	ARMAX 2-2
$\mathcal{C}_1$	0.497	0.391	0.290	0.335
$\mathcal{C}_2$	-1009	-1350	-1784	-1565
$\mathcal{C}_3$	0.499	0.394	0.292	0.340

References

L. Dugard and I.D. Landau. Recursive output error identification algorithms: Theory and evaluation. *Automatica*, pages 443–462, 1980.

L. Dugard and I.D. Landau, editors. *Commande Adaptative des Systèmes. Théorie, Méthodologie, Applications*, ENSIEG, BP 46, 38402 St-Martin d'Hères (France), 1990. Laboratoire d'Automatique de Grenoble.

R. Fletcher. *Practical Methods of Optimization*. Wiley, Chichester, 1991.

U. Forssel and L. Ljung. Closed-loop identification revisited. *Automatica*, 35 (7):1215–1241, 1999.

P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic Press, London, 1981.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. John Hopkins, Baltimore, 1989.

G.C. Goodwin and K.S. Sin. *Adaptive Filtering, Prediction and Control*. Prentice Hall, Englewood Cliffs, 1984.

R. Isermann. *Digital Control Systems*, volume II. Stochastic Control, Multivariable Control, Adaptive Control, Applications. Springer-Verlag, 2nd edition, 1991.

I.D. Landau. *Identification et Commande des Systèmes*. Hermès, Paris, 1988.

I.D. Landau. *System Identification and Control Design*. Prentice Hall, Englewood Cliffs, 1990.

I.D. Landau and A. Besançon Voda, editors. *Identification des Systèmes*. Hermès, Paris, 2001.

I.D. Landau and A. Karimi. Recursive algorithms for identification in closed loop - a unified approach and evaluation. *Automatica*, 33(8):1499–1523, 1997.

L. Ljung. *System Identification. Theory for the User*. Prentice Hall, Englewood Cliffs, 1987.

L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, Massachussetts, 1986.

R.H. Middleton and G.C. Goodwin. *Digital Control and Estimation*. Prentice Hall, Englewood Cliffs, 1990.

T. Söderström and P. Stoica. *System Identification*. Prentice Hall, New York, 1989.

T. Söderström, L. Ljung, and I. Gustavsson. A theoretical analysis of recursive identification methods. *Automatica*, 14:231–244, 1978.

P. Van Den Hof and R. Schrama. Identification and control, closed-loop issues. *Automatica*, 31(12):1751–1770, 1995.

P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic, Dordecht, 1996.

E. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data*. Communications and Control Engineering. Springer-Verlag, London, 1997.

Y. Zhu and T. Backx. *Identification of Multivariable Industrial Processes*. Springer-Verlag, London, 1993.

Part IV

Discrete Time Control

Chapter 13

Digital Control

This chapter mainly concerns pole-placement control in the case of discrete time. By its general character, pole-placement control is very important; indeed, it can cover other specialized types of control such as discrete PID, linear quadratic control studied in Chap. 14 in relation to optimal control, generalized predictive control studied in Chap. 15, model predictive control studied in Chap. 16. Discrete internal model control will also be examined, as well as general characters of adaptive control.

13.1 Pole-Placement Control

13.1.1 Influence of Pole Position

In continuous time, the influence of the position of the poles of the closed-loop transfer function on the overshoot, the rise time and the settling time was studied (Fig. 4.2). Using the correspondence $z = \exp(sT_s)$, it is possible to obtain in discrete time an analogous figure in the unit circle of the complex plane. The drawing in Fig. 13.1 was realized with the following values: $\zeta = 0.7$, $\omega_n = 1$, $T_s = 1$. Notice that the fast poles correspond to $z \approx 1$ with the integrator pole $z = 1$. The results in this Figure can be compared with the step responses in Fig. 9.21 with respect to the pole position.

13.1.2 Control Synthesis by Pole-Placement

Pole-placement control was already studied in the continuous case in Chap. 4. As Aström and Wittenmark (1989) mentions, model reference control is considered as a particular case of pole-placement and is considered in the same framework.

The pole-placement controller or RST controller, is described by the canonical structure in Fig. 13.2. It can be used for unstable or stable systems as well.

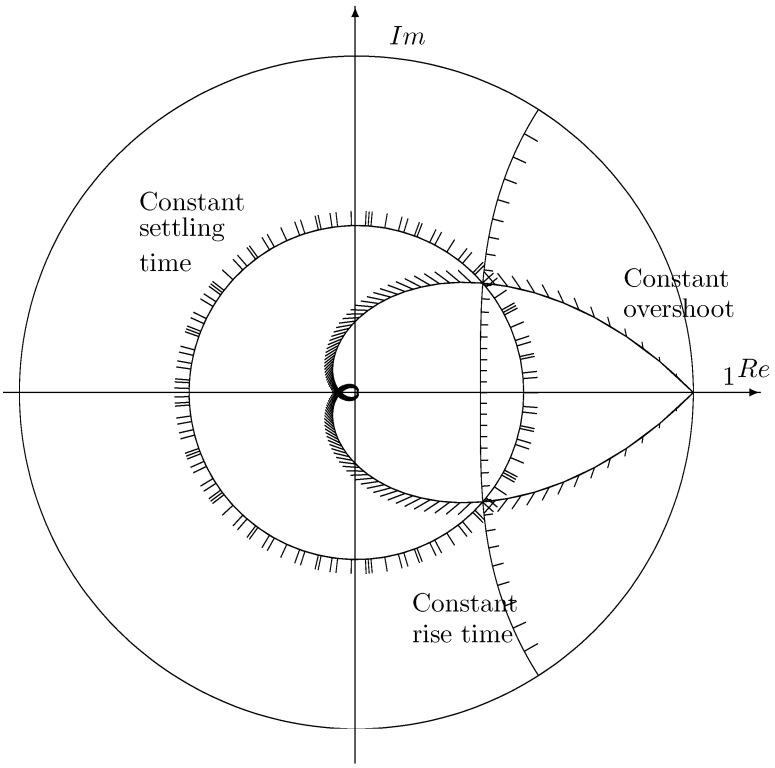


Figure 13.1: Loci of constant overshoot, rise time and settling time in the complex plane in discrete time. The dashed zone is not convenient for the poles

It is possible to present pole-placement with the forward shift operator q (Aström and Wittenmark, 1989) or with the backward shift operator q^{-1} (Landau, 1990), and also with the δ operator (Middleton and Goodwin, 1990). This last choice introduces a better numerical robustness. The choice of operator q or q^{-1} influences, in particular, the solving of the Bezout equation (De Larminat, 1993).

In this introductory part, in order to avoid problems due to the use of the operator q^{-1}, the operator q will be used. Moreover, the parallel with continuous pole-placement (Sect. 4.9) becomes still more obvious.

The process transfer function can be represented as

$$H(q) = \frac{y(t)}{u(t)} = \frac{B(q)}{A(q)} \tag{13.1}$$

with

$$\begin{array}{ll} A(q) & = q^n + a_1 q^{n-1} + \ldots + a_{n-1} q + a_n \\ B(q) & = b_1 q^{n-1} + \ldots + b_{n-1} q + b_n \end{array} \tag{13.2}$$

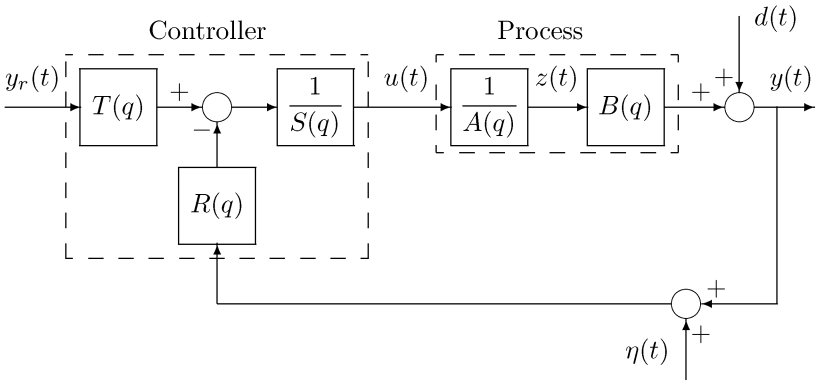


Figure 13.2: Pole-placement or RST control

so that

$$\frac{y(t)}{u(t)}=\frac{b_1q^{n-1}+\ldots+b_{n-1}q+b_n}{q^n+a_1q^{n-1}+\ldots+a_{n-1}q+a_n}=\frac{b_1q^{-1}+\ldots+b_{n-1}q^{-n+1}+b_nq^{-n}}{1+a_1q^{-1}+\ldots+a_{n-1}q^{-n+1}+a_nq^{-n}} \tag{13.3}$$

The transfer function $H(q)$ is strictly proper: $\deg A > \deg B$. The coefficient b_0 is zero to emphasize that the output is always delayed by at least one sampling period with respect to the input. The system is causal. There always exists a delay equal to at least one sampling period; if, moreover, the system presents an effective delay d, the d following coefficients $b_1, b_2, \ldots$ are zero. The polynomial $A(q)$ is monic ($a_0 = 1$). The corresponding difference equation is

$$y(t) + a_1y(t-1) + \ldots + a_ny(t-n) = b_1u(t-1) + \ldots + b_nu(t-n) \tag{13.4}$$

The polynomials $B(q)$ and $A(q)$ are assumed to be coprime.

Note in the same manner the polynomials $R(q)$, $S(q)$, $T(q)$

$$\begin{aligned} R(q) &= r_0q^m + r_1q^{m-1} + \ldots + r_{m-1}q + r_m \\ S(q) &= q^m + s_1q^{m-1} + \ldots + s_{m-1}q + s_m \\ T(q) &= T_0q^m + \ldots + T_{m-1}q + T_m \end{aligned} \tag{13.5}$$

so that the controller is proper: $\deg S \geq \deg R$ and $\deg S \geq \deg T$. According to the block diagram in Fig. 13.2, the control law is

$$u(t) = -\frac{R(q)}{S(q)}\,y(t) + \frac{T(q)}{S(q)}\,y_r(t) \tag{13.6}$$

where $y_r(t)$ is the reference.

The closed-loop output results

$$\begin{aligned} y(t) = \quad & \frac{B(q)T(q)}{A(q)S(q)+B(q)R(q)}\,y_r(t) + \frac{A(q)S(q)}{A(q)S(q)+B(q)R(q)}\,d(t) \\ & -\frac{B(q)R(q)}{A(q)S(q)+B(q)R(q)}\,\eta(t) \end{aligned} \tag{13.7}$$

The pole-placement means that the closed-loop poles are specified, corresponding to the zeros of the polynomial $P(q)$ defined by the Bezout equation

$$P(q) = A(q)S(q) + B(q)R(q) \tag{13.8}$$

with

$$P(q) = P_0 q^p + P_1 q^{p-1} + \ldots + P_{p-1} q + P_p \tag{13.9}$$

From the equality of the degrees of P and the product AS, we deduce $p = n+m$. We reason as if s_0 were unknown. The number of equations corresponds simultaneously to the number of unknown coefficients of polynomials $R(q)$ and $S(q)$, and to the number of specified coefficients of polynomial $P(q)$. So that the solution is unique, we impose $p+1 = 2m+2$, hence $p = 2n-1, m = n-1$, thus for a simply proper controller $\deg P = 2\deg A - 1$ and $\deg S = \deg R = \deg A - 1$.

Given the polynomial $P(q)$ and a proper controller, the solution of the Bezout Eq. (13.8) gives the polynomials $R(q)$ and $S(q)$ by solving the following system

$$\mathcal{S} \begin{bmatrix} s_0 \\ \vdots \\ s_{n-1} \\ -- \\ r_0 \\ \vdots \\ r_{n-1} \end{bmatrix} = \begin{bmatrix} P_0 \\ \vdots \\ P_{n-1} \\ -- \\ P_n \\ \vdots \\ P_{2n-1} \end{bmatrix} \tag{13.10}$$

with

$$\mathcal{S} = \underbrace{\left[\begin{array}{ccccc|} a_0 & 0 & \ldots & & 0 \\ a_1 & a_0 & & & \vdots \\ \vdots & & \ddots & & 0 \\ \vdots & & & & \\ a_{n-1} & a_{n-2} & \ldots & & a_0 \\ -- & -- & -- & -- & -- \\ a_n & a_{n-1} & \ldots & & a_1 \\ 0 & a_n & \ddots & & \vdots \\ \vdots & \ddots & & & \vdots \\ & & & a_n & a_{n-1} \\ 0 & \ldots & & 0 & a_n \end{array}\right.}_{n \text{ columns}} \underbrace{\left.\begin{array}{ccccc} 0 & \ldots & & & 0 \\ b_1 & \ddots & & & \vdots \\ b_2 & \ddots & & & \\ \vdots & & \ddots & & \vdots \\ b_{n-1} & b_{n-2} & \ldots & b_1 & 0 \\ -- & -- & -- & -- & -- \\ b_n & b_{n-1} & \ldots & & b_1 \\ 0 & b_n & \ldots & & b_2 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & & b_n & b_{n-1} \\ 0 & \ldots & & 0 & b_n \end{array}\right]}_{n \text{ columns}} \tag{13.11}$$

where $R(q)$ and $S(q)$ follow Eq. (13.5) with $m = n-1$. In fact, in the previous system $a_0 = 1$, $s_0 = 1$ and $P_0 = 1$. The matrix to be inverted during the solving is a Sylvester matrix $\mathcal{S}$, of dimension $(2n) \times (2n)$, for this proper controller.

So that the Sylvester matrix is nonsingular, thus invertible, it is necessary and sufficient that the polynomials $A(q)$ and $B(q)$ are coprime.

A strictly proper controller ($\deg S = \deg R + 1$) offers better high-frequency behaviour. It must verify $\deg S = \deg A$, $\deg R = \deg A - 1$ and $\deg P = 2\deg A$. In the case of a strictly proper controller, the calculation is performed in a very similar manner with the Sylvester matrix of dimension $(2n+1) \times (2n+1)$, and coefficient $r_0 = 0$ as follows

$$\mathcal{S} \begin{bmatrix} s_0 \\ \vdots \\ s_n \\ -- \\ r_1 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} P_0 \\ \vdots \\ P_n \\ -- \\ P_{n+1} \\ \vdots \\ P_{2n} \end{bmatrix} \tag{13.12}$$

Again, in this system, $a_0 = 1$, $s_0 = 1$ and $P_0 = 1$. $R(q)$ and $S(q)$ are given by

$$\begin{aligned} R(q) &= r_1 q^{n-1} + r_2 q^{n-2} + \ldots + r_{n-1} q + r_n \\ S(q) &= q^n + s_1 q^{n-1} + \ldots + s_{n-1} q + s_n \end{aligned} \tag{13.13}$$

with

$$\mathcal{S} = \underbrace{\left[\begin{array}{ccccccc} a_0 & 0 & \ldots & & & 0 & | \\ a_1 & a_0 & & & & \vdots & | \\ \vdots & & \ddots & & & 0 & | \\ \vdots & & & & & & | \\ a_{n-1} & \ldots & & & a_0 & 0 & | \\ a_n & a_{n-1} & \ldots & & & a_0 & | \\ -- & -- & -- & -- & -- & -- & | \\ 0 & a_n & a_{n-1} & \ldots & & a_1 & | \\ \vdots & \ddots & a_n & \ldots & & \vdots & | \\ \vdots & & \ddots & \ddots & & \vdots & | \\ & & & & a_n & a_{n-1} & | \\ 0 & \ldots & & & 0 & a_n & | \end{array}\right]}_{n+1 \text{ columns}} \underbrace{\left[\begin{array}{ccccc} 0 & \ldots & & & 0 \\ 0 & \ldots & & & \vdots \\ b_1 & 0 & & & \\ \vdots & \ddots & \ddots & & \vdots \\ b_{n-2} & \ldots & b_1 & 0 & 0 \\ b_{n-1} & b_{n-2} & \ldots & b_1 & 0 \\ -- & -- & -- & -- & -- \\ b_n & b_{n-1} & \ldots & & b_1 \\ 0 & b_n & \ldots & & b_2 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & & b_n & b_{n-1} \\ 0 & \ldots & & 0 & b_n \end{array}\right]}_{n \text{ columns}} \tag{13.14}$$

It is more customary to use the polynomials with the operator q^{-1} rather than q; note that Aström and Wittenmark (1989) and Middleton and Goodwin (1990) use the operator q and that De Larminat (1993) even recommends its use, although it is not causal (as well as the Laplace variable s). For a transfer function whose numerator and denominator are formally polynomials of equal degree in q and q^{-1}, the representation is equivalent.

If we had used the polynomials with the backward shift operator q^{-1}, we would have found a solution to the Bezout equation

$$P(q^{-1}) = A(q^{-1})S(q^{-1}) + B(q^{-1})R(q^{-1}) \tag{13.15}$$

for a polynomial $P(q^{-1})$ of any degree, which is not the case for Eq. (13.8). In fact, it is not desirable that $P(q^{-1})$ is of any degree, as a too low degree of $P(q^{-1})$ with respect to the previous conditions leads to implicit placement of poles at the origin. If on one hand the degrees of $A(q^{-1})$ and of $B(q^{-1})$ are identical and on the other hand those of $S(q^{-1})$ and $R(q^{-1})$ are also identical, the solving of the Bezout equation in q or in q^{-1} is equivalent.

In the case where the operator q^{-1} is used, the system is modelled by

$$H(q^{-1}) = \frac{y(t)}{u(t)} = \frac{B(q^{-1})}{A(q^{-1})} = q^{-d-1}\,\frac{B'(q^{-1})}{A(q^{-1})} \tag{13.16}$$

with the delay d and polynomials

$$\begin{array}{ll} A(q^{-1}) & = 1 + a_1 q^{-1} + \ldots + a_{n_a-1} q^{-n_a+1} + a_{n_a} q^{-n_a} \\ B'(q^{-1}) & = b'_0 + \ldots + b'_{n'_b-1} q^{-n'_b+1} + b'_{n'_b} q^{-n'_b} \end{array} \tag{13.17}$$

The control law is

$$S(q^{-1})u(t) = T(q^{-1})y_r(t) - R(q^{-1})y(t) \tag{13.18}$$

where the polynomials $R(q^{-1})$ and $S(q^{-1})$ are obtained by solving the Bezout equation

$$P(q^{-1}) = A(q^{-1})\,S(q^{-1}) + q^{-d-1}\,B'(q^{-1})\,R(q^{-1}) \tag{13.19}$$

and the closed-loop output is equal to

$$y(t)=\frac{q^{-d-1}\,B'(q^{-1})\,T(q^{-1})}{A(q^{-1})\,S(q^{-1}) + q^{-d-1}\,B'(q^{-1})\,R(q^{-1})}\,y_r(t)=\frac{q^{-d-1}\,B'(q^{-1})\,T(q^{-1})}{P(q^{-1})}\,y_r(t) \tag{13.20}$$

Example 13.1: **Comparison Between q and q^{-1} Models**

Consider the transfer function expressed by q

$$\frac{y(t)}{u(t)} = H(q) = \frac{B(q)}{A(q)} = \frac{0q^2 + 0q + 0.5}{q^2 - 0.5q + 0.3} \tag{13.21}$$

Formally, $A(q)$ and $B(q)$ have same degree 2. However, the first two coefficients of $B(q)$ are zero, thus the delay of $y(t)$ with respect to $u(t)$ is two sampling periods, including the normal delay of one sampling period of the output with respect to the input. The following transfer function in q^{-1} is equivalent

$$\begin{array}{ll} \dfrac{y(t)}{u(t)} & = H(q^{-1}) = \dfrac{B(q^{-1})}{A(q^{-1})} = \dfrac{0 + 0q^{-1} + 0.5q^{-2}}{1 - 0.5q^{-1} + 0.3q^{-2}} \\ & = q^{-2}\,\dfrac{0.5}{1 - 0.5q^{-1} + 0.3q^{-2}} \end{array} \tag{13.22}$$

where, besides the normal delay, there exists an effective delay $d = 1$ of one period.

In a general way, if we want to use the same coefficients in polynomials A and B whatever they are in q or in q^{-1}, they must be written as

$$y(t) = \frac{B(q^{-1})}{A(q^{-1})}u(t) \Longleftrightarrow y(t) = \frac{B(q)}{q^{nb}}\,\frac{q^{na}}{A(q)}u(t) = q^{na-nb}\frac{B(q)}{A(q)}u(t) \qquad (13.23)$$

where n_a is the degree of polynomial A, n_b of polynomial B.

Regulation Behaviour

In order to guarantee a zero steady-state error, in general, an integrator is included in the controller so that the polynomial $S(q)$ is replaced by $H_1(q)S(q)$ with $H_1(q) = q - 1$. The addition of an integrator in the controller aims to reject the constant disturbances in agreement with the internal model principle (Sect. 5.9).

On the other hand, a robustness filter is also added in the feedback loop (Fig. 13.3), transforming polynomial $R(q)$ into $H_2(q)R(q)$ with

$$H_2(q) = \frac{q-\alpha}{1-\alpha} \quad ; \qquad 0 < \alpha \ll 1 \qquad (13.24)$$

In these conditions, the Bezout equation becomes

$$P(q) = A(q)\,H_1(q)\,S(q) + B(q)\,H_2(q)\,R(q) \qquad (13.25)$$

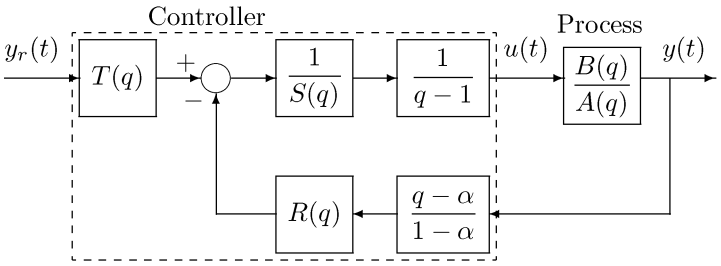


Figure 13.3: Pole-placement control with integrator and robustness filter

Tracking Behaviour

The reference trajectory $y_{ref}(t)$ is either the set point $y_r(t)$ or defined with respect to a reference model (Fig. 13.4)

$$H_m(q) = \frac{B_m(q)}{A_m(q)} \qquad (13.26)$$

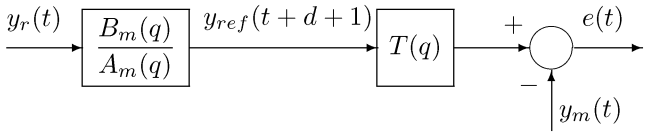


Figure 13.4: Reference trajectory and set point

so that

$$y_{ref}(t) = q^{-d-1}\,\frac{B_m(q)}{A_m(q)}y_r(t) \Longleftrightarrow y_{ref}(t+d+1) = \frac{B_m(q)}{A_m(q)}y_r(t) \tag{13.27}$$

in order to take into account the delay $d+1$, with the polynomials

$$\begin{array}{ll} A_m(q) & = q^{nm} + a_{m1}q^{nm-1} + \ldots + a_{nm} \\ B_m(q) & = b_{m0}q^{nm} + b_{m1}q^{nm-1} + \ldots + b_{nm} \end{array} \tag{13.28}$$

The polynomials A_m and B_m can be obtained from a continuous transfer function, e.g. derived from the ITAE criterion, then discretize, for example, by a zero-order holder or the Tustin transformation.

If we separate the polynomial $T(q)$ from the reference model B_m/A_m as in Fig. 13.4, to guarantee a unit gain from Eq. (13.7), we impose

$$T(z) = \left\{ \begin{array}{ll} P(z)/B(1) & \text{if: } B(1) \neq 0 \\ P(z) & \text{if: } B(1) = 0 \end{array} \right. \tag{13.29}$$

Aström and Wittenmark (1989) integrate the reference model B_m/A_m in the polynomial $T(q)$; in these conditions, for the gain to be 1, we need

$$\left\{ \frac{B(z)\,T(z)\,B_m(z)}{P(z)\,A_m(z)} \right\}_{z=1} = 1 \tag{13.30}$$

13.1.3 Relation between Pole-Placement and State Feedback

In this section, where we insist on causal relations and where the objective is to establish an analogy between pole-placement and state feedback, the operator q^{-1} is used. According to Fig. 13.2, we introduce the partial state $z(t)$ as

$$A(q^{-1})z(t) = u(t) \quad ; \qquad y(t) = B(q^{-1})z(t) \tag{13.31}$$

It is possible to factorize the pole-placement polynomial $P(q^{-1})$ as

$$P(q^{-1}) = A_o(q^{-1})[A(q^{-1}) + K(q^{-1})] \tag{13.32}$$

with the polynomial $K(q^{-1})$ being of degree n such that

$$K(q^{-1}) = k_1 q^{-1} + \ldots + k_{n-1}q^{-n+1} + k_n q^{-n} \tag{13.33}$$

and $A_o(q^{-1})$ being of degree $n-1$. We then define a polynomial $P_2(q^{-1})$ such that

$$P_2(q^{-1}) = S(q^{-1}) - A_o(q^{-1}) \tag{13.34}$$

Using Eqs. (13.18) and (13.27), we deduce

$$[A_o(q^{-1}) + P_2(q^{-1})]u(t) = T(q^{-1})y_r(t+d+1) - R(q^{-1})y(t) \tag{13.35}$$

By introducing the partial state $z(t)$, we have

$$A_o(q^{-1})u(t) = T(q^{-1})y_r(t+d+1) - [P_2(q^{-1})A(q^{-1}) + R(q^{-1})B(q^{-1})]z(t) \tag{13.36}$$

As

$$\begin{aligned} P_2(q^{-1})A(q^{-1}) + R(q^{-1})B(q^{-1}) &= [S(q^{-1}) - A_o(q^{-1})]A(q^{-1}) + R(q^{-1})B(q^{-1}) \\ &= P(q^{-1}) - A_o(q^{-1})A(q^{-1}) \\ &= A_o(q^{-1})K(q^{-1}) \end{aligned} \tag{13.37}$$

we obtain

$$\begin{aligned} [P_2(q^{-1})A(q^{-1}) + R(q^{-1})B(q^{-1})]z(t) &= A_o(q^{-1})K(q^{-1})z(t) \\ &= P_2(q^{-1})u(t) + R(q^{-1})y(t) \end{aligned} \tag{13.38}$$

and

$$\begin{aligned} &A_o(q^{-1})u(t) = T(q^{-1})y_r(t+d+1) - A_o(q^{-1})K(q^{-1})z(t) \quad \Longrightarrow \\ &u(t) = \frac{T(q^{-1})}{A_o(q^{-1})}y_r(t+d+1) - K(q^{-1})z(t) \end{aligned} \tag{13.39}$$

Equations (13.38) and (13.39) correspond to the scheme in Fig. 13.5. As the polynomial $A_o(q^{-1})$ allows us to estimate the partial state vector $z(t)$ according to Eq. (13.38), it is called the observer polynomial and specifies the observer dynamics.

The closed-loop equation of the system can then be formulated as

$$\begin{cases} A_o(q^{-1})[A(q^{-1}) + K(q^{-1})]z(t) = T(q^{-1})y_r(t+d+1) \\ y(t) = B(q^{-1})z(t) \end{cases} \tag{13.40}$$

Indeed, we verify that the denominator of the closed-loop transfer function is equal to the pole-placement polynomial

$$A_o(q^{-1})[A(q^{-1}) + K(q^{-1})] = P(q^{-1}) \tag{13.41}$$

The system can also be expressed in state-space equations. By setting the state vector composed by the partial states $z(t)$, according to Eq. (13.31)

$$x(t) = [z(t-1)\dots z(t-n)]^T \tag{13.42}$$

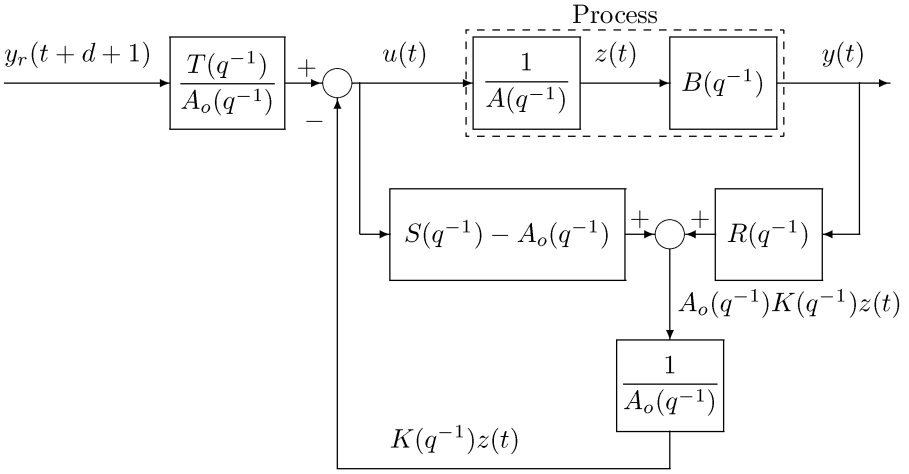


Figure 13.5: Pole-placement representation with partial state $z(t)$ and the observer polynomial $A_o(q^{-1})$

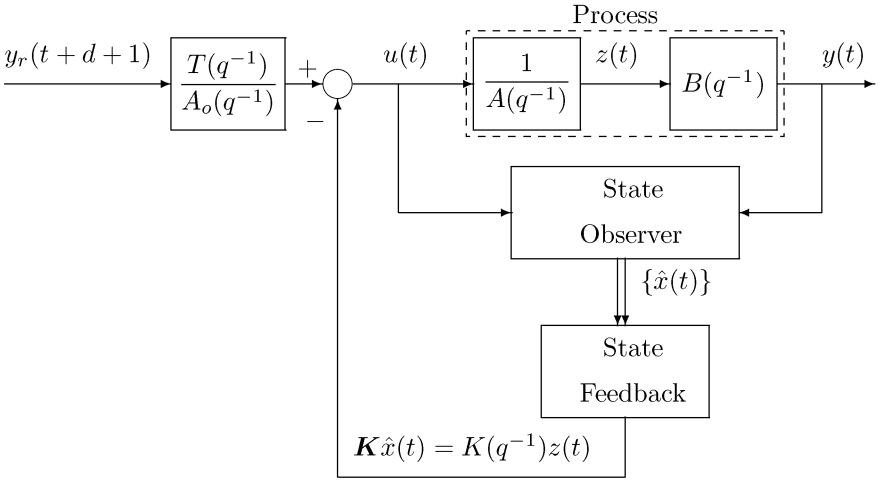


Figure 13.6: State-space representation of pole-placement with partial state $z(t)$ and the observer polynomial $A_o(q^{-1})$

the model is written in the controllable form as

$$x(t+1) = \begin{bmatrix} -a_1 & \dots & & -a_n \\ 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & & 0 \quad 1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} u(t) \qquad (13.43)$$

$$y(t) = [b_1 \dots b_n]\, x(t)$$

By setting $\boldsymbol{K}$ as a state feedback gain vector equal to

$$\boldsymbol{K} = [k_1 \ \ldots \ k_n] \tag{13.44}$$

Eq. (13.38) can be seen in the form

$$A_o(q^{-1})K(q^{-1})z(t) = A_o(q^{-1})\boldsymbol{K}x(t) = P_2(q^{-1})u(t) + R(q^{-1})y(t) \tag{13.45}$$

assuming that all the states $x(t)$ are measurable. The equation expressing the state feedback (13.39) is then

$$u(t) = \frac{T(q^{-1})}{A_o(q^{-1})} y_r(t+d+1) - \boldsymbol{K}x(t) \tag{13.46}$$

By using the state feedback (13.46), the closed-loop model becomes

$$\begin{aligned} x(t+1) &= \begin{bmatrix} -a_1 - k_1 & \ldots & & -a_n - k_n \\ 1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & & 0 \quad 1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} y_r^*(t+d+1) \\ y(t) &= [b_1 \ldots b_n]\, x(t) \end{aligned} \tag{13.47}$$

by setting

$$y_r^*(t+d+1) = \frac{T(q^{-1})}{A_o(q^{-1})} y_r(t+d+1) \tag{13.48}$$

We deduce the characteristic polynomial

$$z^n + (a_1 + k_1)z^{n-1} + \ldots + (a_n + k_n) \tag{13.49}$$

which shows that the closed-loop poles can be specified by means of coefficients k_i.

In the common case where the states $x(t)$ are not all measurable, it is necessary to use an observer to estimate the states $\hat{x}(t)$ (Goodwin and Sin, 1984; Vidyasagar, 1985). The system is described in state space in the observable form and with gain $\boldsymbol{L}$ of the observer

$$\begin{aligned} \hat{x}_o(t+1) = &\begin{bmatrix} -a_1 & 1 & 0 & \ldots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ldots & \ddots & 1 \\ -a_n & 0 & \ldots & \ldots & 0 \end{bmatrix} \hat{x}_o(t) + \begin{bmatrix} b_1 \\ \vdots \\ \\ \vdots \\ b_n \end{bmatrix} u(t) \\ &+ \begin{bmatrix} l_1 \\ \vdots \\ \\ \vdots \\ l_n \end{bmatrix} (y(t) - [10 \ldots 0]\hat{x}_o(t)) \end{aligned} \tag{13.50}$$

It is possible to go from the states x to the states x_o by a transformation matrix $\boldsymbol{P}$: $x_o(t) = \boldsymbol{P}x(t)$, so that we obtain the estimation of the states

$$\hat{x}(t+1) = \boldsymbol{P}^{-1} \begin{bmatrix} -a_1 - l_1 & 1 & 0 & \dots & 0 \\ \vdots & & 0 & \ddots & \vdots \\ \vdots & & \vdots & \ddots & 0 \\ \vdots & & \vdots & \ddots & 1 \\ -a_n - l_n & 0 & \dots & \dots & 0 \end{bmatrix} \boldsymbol{P}\,\hat{x}(t) + \boldsymbol{P}^{-1} \begin{bmatrix} b_1 \\ \vdots \\ \vdots \\ b_n \end{bmatrix} u(t) + \boldsymbol{P}^{-1} \begin{bmatrix} l_1 \\ \vdots \\ \vdots \\ l_n \end{bmatrix} y(t) \tag{13.51}$$

In Eq. (13.46), the states $x(t)$ are at last replaced by their estimation $\hat{x}(t)$. The observer and the state feedback are displayed in Fig. 13.6.

13.1.4 General Pole-Placement Design

The choice of polynomial $P(q)$ determines the type of pole-placement obtained. The polynomial $P(z)$ can thus be chosen by spectral factorization (Aström and Wittenmark, 1989) from a quadratic criterion

$$J = \int_0^\infty y(t)^2 + \rho u(t)^2 dt \tag{13.52}$$

from which we draw

$$\rho\, A(z)\, A^*(z) + z^d\, B(z)\, B^*(z) = \rho\, P(z)\, P^*(z) \tag{13.53}$$

where $A^*(z)$ is the polynomial obtained from $A(z)$ by replacing z by $1/z$. In this case, the parameter ρ alone is sufficient to specify $P(z)$.

It is also possible to follow the same procedure as for continuous pole-placement (Aström and Wittenmark, 1989); the presentation is realized in a general framework before examining particular cases.

On one hand, if we want to take into account the reference model and if we integrate the model $B_m(q)/A_m(q)$ in $T(q)$, it is necessary that the polynomial $A_m(q)$ is a divider of $P(q)$.

According to Eq. (13.7), the process zeros which are those of $B(q)$ are also those of the closed-loop transfer function, except if common roots exist between $B(q)$ and $P(q)$. Unstable or weakly damped zeros cannot be eliminated. The polynomial $B(q)$ is then factorized as

$$B(q) = B^+(q)\, B^-(q) \tag{13.54}$$

where $B^+(q)$ is a polynomial containing all stable and well-damped zeros, which can therefore be eliminated, and $B^-(q)$ is a polynomial containing all other

zeros. $B^+(q)$ is chosen to be monic to make the factorization unique (Aström, 1980).

Lastly, the polynomial $P(q)$ is factorized as

$$P(q) = A_m(q)\, B^+(q)\, A_o(q) \tag{13.55}$$

where $A_o(q)$ is the polynomial giving the observer dynamics.

The Bezout equation is then written in the general form

$$A(q)\, S(q) + B(q)\, R(q) = A_m(q)\, B^+(q)\, A_o(q) \tag{13.56}$$

We deduce that $B^+(q)$ must be a divider of $S(q)$, so that we can set

$$S(q) = S'(q)\, B^+(q) \tag{13.57}$$

from which results the simplified Bezout equation which gives $R(q)$ and $S'(q)$

$$A(q)\, S'(q) + B^-(q)\, R(q) = A_m(q)\, A_o(q) \tag{13.58}$$

and the expression of the closed-loop output

$$y(t) = \frac{B^-(q)\, T(q)}{A_m(q)\, A_o(q)}\, y_r(t) \tag{13.59}$$

In order to have the same output as that specified by the reference model, it is necessary that

$$\frac{B^-(q)\, T(q)}{A_m(q)\, A_o(q)} = \frac{B_m(q)}{A_m(q)} \Longrightarrow B^-(q)\, T(q) = A_o(q)\, B_m(q) \tag{13.60}$$

Rigorously, this equation implies model tracking, as the poles and the zeros are considered simultaneously, while the strict pole-placement only takes into account the poles. Aström and Wittenmark (1989) insist on the importance in considering the zeros in order to realize a good control system. Moreover, they advise not to change the open-loop zeros of the process, but to keep them in the model of the tracking dynamics $B_m(q)/A_m(q)$.

From Eq. (13.60), we deduce that $B^-(q)$ must divide $B_m(q)$, thus by setting

$$B_m(q) = B^-(q)\, B'_m(q) \tag{13.61}$$

we obtain the polynomial $T(q)$

$$T(q) = A_o(q)\, B'_m(q) = \frac{A_o(q)\, B_m(q)}{B^-(q)} \tag{13.62}$$

Landau (1990), using the backward shift operator q^{-1}, proposes a less severe condition for Eq. (13.60) by only demanding that the gain between output and set point is unitary, and that the output dynamics is that of polynomial

$A_m(q^{-1})$. In these conditions, it suffices that the polynomial $T(q^{-1})$ is equal to

$$T(q^{-1}) = \begin{cases} \dfrac{A_o(q^{-1})}{B'(1)} & \text{if:} \quad B'(1) \neq 0 \\ A_o(q^{-1}) & \text{if:} \quad B'(1) = 0 \end{cases} \tag{13.63}$$

This approach is, of course, usable with the forward shift operator q.

The degree conditions are

$$\begin{array}{l} \deg P \geq 2\deg A - 1 \Longrightarrow \deg A_o \geq 2\deg A - \deg A_m - \deg B^+ - 1 \\ \deg A_m - \deg B_m \geq \deg A - \deg B \end{array} \tag{13.64}$$

Then, it is possible to look at the particular pole-placement cases. In all cases, the control law will be given by Eq. (13.18). The reference model B_m/A_m must be specified by the user and must satisfy the degree condition.

Pole-Placement without Zero Compensation

In the absence of zero compensation, it suffices to take $B^+(q) = 1$ by comparison with the general case. We obtain the Bezout equation, which gives $R(q)$ and $S(q)$

$$A(q)\,S(q) + B(q)\,R(q) = A_m(q)\,A_o(q) \tag{13.65}$$

and

$$T(q) = \frac{A_o(q)\,B_m(q)}{B(q)} \tag{13.66}$$

Pole-Placement with all Zeros Compensated

When all zeros (necessarily stable) are compensated, it suffices to take $B^+(q) = B(q)/b_0$ by comparison with the general case, so that $B^+(q)$ is monic. We obtain the Bezout equation, which gives $R(q)$ and $S'(q)$

$$A(q)\,S'(q) + b_0\,R(q) = A_m(q)\,A_o(q) \tag{13.67}$$

with

$$S(q) = S'(q)\,B^+(q) \tag{13.68}$$

and

$$T(q) = A_o(q)\,B_m(q)/b_0 \tag{13.69}$$

Pole-Placement with Stable and Well-Damped Zeros Compensated

This case is the most general, and thus resumes the equations previously developed in the general framework.

$B(q)$ is factorized with $B^+(q)$ a monic polynomial containing the stable and well-damped zeros.

$$B(q) = B^+(q)\,B^-(q) \tag{13.70}$$

The polynomial $P(q)$ is equal to

$$P(q) = A_m(q)\,B^+(q)\,A_o(q) \tag{13.71}$$

where $A_o(q)$ specifies the observer dynamics.

The simplified Bezout equation

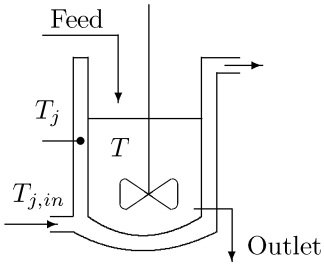
$$A(q)\,S'(q) + B^-(q)\,R(q) = A_m(q)\,A_o(q) \tag{13.72}$$

gives $R(q)$ and $S'(q)$ such that: $S(q) = S'(q)\,B^+(q)$.

The polynomial $T(q)$ is equal to

$$T(q) = \frac{A_o(q)\,B_m(q)}{B^-(q)} \tag{13.73}$$

Example 13.2: **Pole-Placement Control of a Chemical Reactor**



The pole-placement control will be applied to the chemical reactor described in Sect. 19.2.

The model of the pilot chemical reactor (Sect. 19.2) obtained by identification by means of recursive extended least squares is a stable system with transfer function

$$H(q) = \frac{B(q)}{A(q)} = \frac{1.203\,q + 0.426}{q^2 - 1.083\,q + 0.104}$$

The total time delay $d + 1$ is then equal to 1. This transfer function presents two stable poles at $z = 0.976$ and $z = 0.106$ and a stable zero at $z = -0.354$. In the present case, the sampling period is $T_s = 5\,\mathrm{s}$.

The observer is chosen so as to ensure faster dynamics than the process one. A first-order continuous filter with time constant $\tau = 4\,\mathrm{s}$ is taken and discretized, hence

$$A_o(q) = q - 0.286$$

a) First, a pole-placement without zero compensation is realized.
An integrator $H_1 = q - 1$ is added to the polynomial $S(q)$ in order to reject step-like disturbances, and a robustness filter $H_2 = (q - 0.8)/(1 - 0.8)$ is added to the polynomial $R(q)$. The degree conditions lead to choose a polynomial $P(q)$ of degree 5. The polynomials $R(q)$ and $S(q)$, which will be calculated by the Bezout equation, will both have degree 2.

As a reference model, a transfer function optimal with respect to the ITAE criterion (Table 4.1) is chosen

$$G_m(s) = \frac{\omega_0^4}{s^4 + 2.1\omega_0 s^3 + 3.4\omega_0^2 s^2 + 2.7\omega_0^3 + \omega_0^4}$$

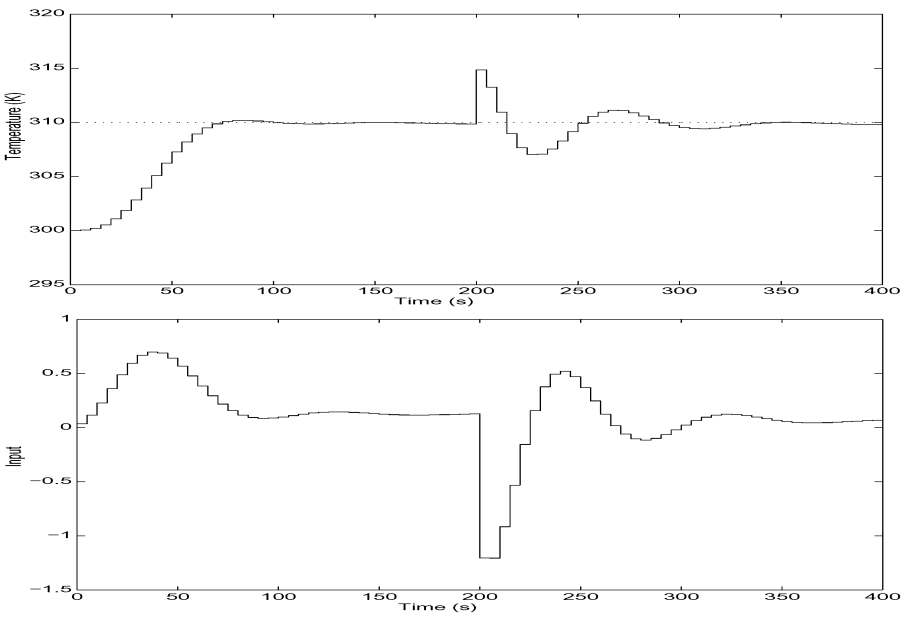


Figure 13.7: Pole-placement control without zero compensation of the identified linear model of the chemical reactor: output temperature (top), variations of the input (bottom). Set point step of amplitude 10 K at $t = 0$ and step disturbance of amplitude 5 K at $t = 40T_s$ at the output

with $\omega_0 = 0.3/T_s$. It is then discretized with a zero-order holder, thus

$$\frac{B_m(q)}{A_m(q)} = \frac{10^{-3}(0.296q^3 + 2.848q^2 + 2.511q + 0.203)}{q^4 - 3.281q^3 + 4.153q^2 - 2.399q + 0.533}$$

which has the poles $0.818 \pm 0.326i$ and $0.822 \pm 0.103i$.

The polynomial $P(q)$ is equal to

$$P(q) = A_m(q)\, A_o(q) = (q^4 - 3.281q^3 + 4.153q^2 - 2.399q + 0.533)\,(q - 0.286)$$

The Bezout equation is written as

$$A(q)\, H_1(q)\, S(q) + B(q)\, H_2(q)\, R(q) = A_m(q)\, A_o(q) \quad \Longrightarrow$$
$$(q^2 - 1.083\, q + 0.104)\,(q - 1)\, S(q) + (1.203\, q + 0.426)\,(q - 0.8)/(1 - 0.8)\, R(q) =$$
$$(q^4 - 3.281q^3 + 4.153q^2 - 2.399q + 0.533)\,(q - 0.286)$$

which gives the controller polynomials $R(q)$ and $S(q)$ (not taking into account the integrator and the robustness filter)

$$R(q) = 0.0533q^2 - 0.0897q + 0.039 \quad ; \qquad S(q) = q^2 - 1.805q + 0.829$$

So that $B(q)$ does not divide $B_m(q)$, the previous polynomial $B_m(q)$ is not kept, but is replaced by

$$B_m(z) = \frac{B(z)A_m(1)}{B(1)}$$

We deduce the polynomial of the precompensator

$$T(z) = \frac{A_o(z)\,A_m(1)}{B(1)} = \frac{(z-0.286)0.00586}{1.629} = 0.0036z - 0.00103$$

The system is subjected to a step set point of amplitude 10 K occurring at instant $t = 0$ and converges towards the set point (Fig. 13.7) without noticeable overshoot. Then, after 40 sampling periods, it is subjected to a step disturbance of amplitude 5 K acting at the output, which is perfectly rejected due to the integrator effect introduced in the controller. The input variations are not too violent and are quite acceptable (Fig. 13.7).

b) Then, a second pole-placement with zero compensation is performed.

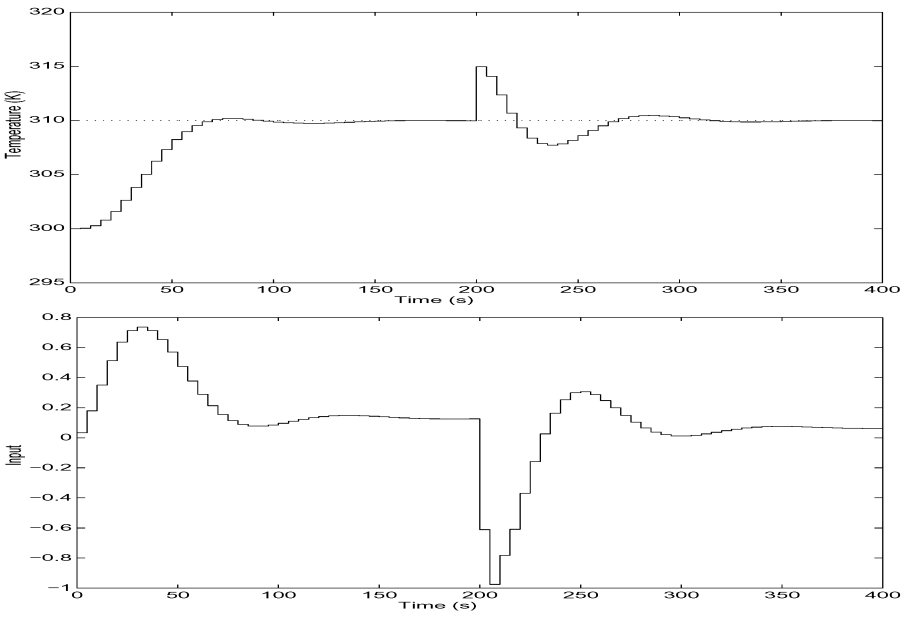


Figure 13.8: Pole-placement control with zero compensation of the identified linear model of the chemical reactor: output temperature (top), variations of the input (bottom). Set point step of amplitude 10 K at $t = 0$ and step disturbance of amplitude 5 K at $t = 40T_s$ at the output

We wish to include an integrator polynomial $H_1(q) = q - 1$ in the controller polynomial $S(q)$ in order to reject step-like disturbances.

The reference model is calculated from an optimal transfer function with regard to the ITAE criterion (Table 4.1)

$$G_m(s) = \frac{\omega_0^3}{s^3 + 1,75\omega_0 s^2 + 2,15\omega_0^2 s + \omega_0^3}$$

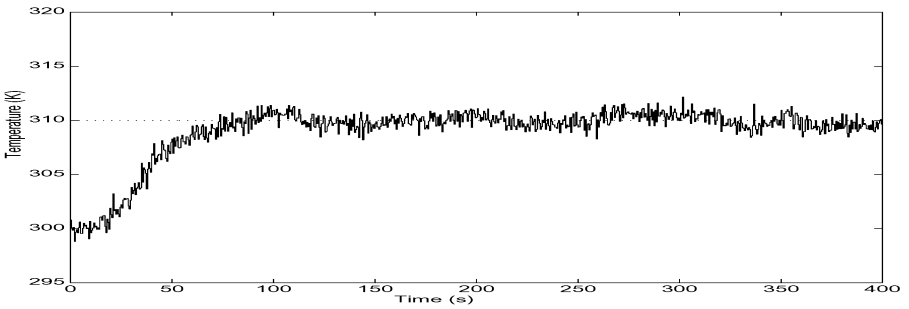


Figure 13.9: Pole-placement control with zero compensation of the identified linear model of the chemical reactor: output temperature. Set point step of amplitude 10 K at $t = 0$. Random disturbance of standard deviation 0.5 K at the output

still keeping $\omega_0 = 0.3/T_s$ in order to really compare the results with those of the previous case. Then, we discretize with a zero-order holder, hence

$$\frac{B_m(q)}{A_m(q)} = \frac{10^{-3}(3.929q^2 + 13.720q + 3.022)}{q^3 - 2.432q^2 + 2.044q - 0.592}$$

which has the poles $0.811 \pm 0.269i$ and 0.809.

The polynomial $B(q)$ is factorized into

$$\begin{aligned} &B(q) = B^+(q)\, B^-(q) \\ &B^+(q) = q + 0.354 \quad ; \quad B^-(q) = 1.203 \end{aligned}$$

where $B^+(q)$ is monic and contains the stable zero. By keeping the same polynomial observer as in part a), the polynomial $P(q)$ is equal to

$$\begin{aligned} P(q) &= A_m(q)\, B^+(q)\, A_o(q) \\ &= (q^3 - 2.432q^2 + 2.044q - 0.592)\,(q + 0.354)\,(q - 0.286) \end{aligned}$$

The simplified Bezout equation is written as

$$\begin{aligned} &A(q)\, H_1(q)\, S'(q) + B^-(q)\, H_2(q)\, R(q) = A_m(q)\, A_o(q) \\ &(q^2 - 1.083q + 0.104)\,(q - 1)\, S'(q) + 1.203\,(q - 0.8)/(1 - 0.8)\, R(q) = \\ &(q^3 - 2.432q^2 + 2.044q - 0.592)\,(q - 0.286) \end{aligned}$$

giving the controller polynomials (not including the integrator and the robustness filter)

$$\begin{aligned} S'(q) &= q - 0.813 \Longrightarrow \\ S(q) &= B^+(q)\, S'(q) \\ &= (q + 0.354)\,(q - 0.813) \\ &= q^2 - 0.459q - 0.288 \\ R(q) &= 0.0295q^2 + 0.000403q - 0.0176 \end{aligned}$$

The polynomial of the precompensator is equal to

$$T(q) = \frac{A_o(q)\,B_m(q)}{B^-(q)} = \frac{(q-0.286)\,10^{-3}(3.929q^2+13.720q+3.022)}{1.203}$$
$$= 10^{-3}(3.266q^3+10.469q^2-0.756q-0.720)$$

The system was subjected to the same step set point of amplitude 10 K as in case a). It converges towards the set point a little faster and without overshoot (Fig. 13.8). Then, after 40 sampling periods, it is subjected to the same disturbance as in a) acting at the output; it is a little better rejected than in case a). The input is slightly smoother (Fig. 13.8). When a random noise of standard deviation 0.5 K affects the output (Fig. 13.9), the system nevertheless maintains the desired behaviour around the set point identical to Figure 13.8.

13.1.5 Digital PID Controller

Discretized PI Analog Controller viewed as an RST

The transfer function of an analog PI controller is

$$G(s) = K\left[1+\frac{1}{\tau_I s}\right] \tag{13.74}$$

Using the backward difference approximation as a discretization operator (T_s being the sampling period)

$$\frac{T_s}{q-1} \tag{13.75}$$

we obtain the following equation of the discretized PI controller

$$(q-1)u(t) = K\left[q-1+\frac{T_s}{\tau_I}\right](y_r(t)-y(t)) \tag{13.76}$$

This equation can be compared to the general equation of the pole-placement (RST) controller (Fig. 13.2)

$$S(q)u(t) = -R(q)\,y(t) + T(q)\,y_r(t) \tag{13.77}$$

By identification, we deduce the polynomials R, S, T which characterize a PI controller

$$\begin{aligned} S(q) &= q-1 \\ R(q) &= -K\left[q-1+\frac{T_s}{\tau_I}\right] \\ T(q) &= -R(q) \end{aligned} \tag{13.78}$$

We notice that $S(q)$ corresponds to a simple integrator. The coefficients of the controller polynomials will be calculated according to the polynomial P, which define the closed-loop behaviour by specifying the values of the poles

$$A(q)\,S(q) + B(q)\,R(q) = P(q) \tag{13.79}$$

The polynomial P can be chosen by proceeding to the discretization of one of the continuous polynomials optimal with regard to the ITAE criterion (Table 4.1).

In the case where the identified process model is limited to polynomials A and B of order 1, the calculation of the coefficients becomes very simple.

Discretized PID Analog Controller seen as an RST

The transfer function of a real analog PID controller is

$$G(s) = K\left[1 + \frac{1}{\tau_I\, s} + \frac{\tau_D\, s}{1 + \frac{\tau_D}{N}\, s}\right] \tag{13.80}$$

with τ_D/N representing a filtering factor of the derivative (derivative time constant τ_D) and τ_I integral time constant.

We proceed with the same operator as previously; s (derivative action) is approximated by $(q-1)/T_s$ and $1/s$ (integral action) is approximated by $T_s/(q-1)$. We deduce the discrete transfer function of the discretized analog PID

$$H(q) = K\left[1 + \frac{T_s}{\tau_I}\frac{1}{q-1} + \frac{N\,(q-1)}{q + \dfrac{T_s N - \tau_D}{\tau_D}}\right] \tag{13.81}$$

By identifying

$$H(q) = \frac{R(q)}{S(q)} \tag{13.82}$$

and imposing an integrator in the monic polynomial $S(q) = (q-1)S'(q)$, which allows the convergence towards the steady-state, we have

$$S(q) = (q-1)\,(q+s_1) \qquad \text{with: } s_1 = \frac{T_s N - \tau_D}{\tau_D} \tag{13.83}$$

It results that

$$\begin{aligned} R(q) &= r_0 q^2 + r_1 q + r_2 \\ r_0 &= K(1+N) \\ r_1 &= K(-1 + s_1 + \frac{T_s}{\tau_I} - 2N] \\ r_2 &= K(s_1(\frac{T_s}{\tau_I} - 1) + N) \end{aligned} \tag{13.84}$$

The coefficients of $R(q)$ and $S(q)$ will be calculated in the same way as for the PI controller by specifying the polynomial $P(q)$ such that

$$A(q)\,S(q) + B(q)\,R(q) = P(q) \tag{13.85}$$

which defines the closed-loop poles whose transfer function is equal to

$$H_{cl}(q) = \frac{y_r(t)}{y(t)} = \frac{B(q)\,T(q)}{A(q)\,S(q) + B(q)\,R(q)} = \frac{B(q)\,T(q)}{P(q)} \tag{13.86}$$

In order to ensure the asymptotic convergence of the output towards the set point, let $H_{cl}(1) = 1$, the tracking polynomial $T(q)$ is chosen as

$$T(z) = \frac{P(1)}{B(1)} = R(1) \tag{13.87}$$

noting that $S(1) = 0$ (because of the integrator term). The closed-loop transfer function thus becomes

$$H_{cl}(z) = \frac{P(1)}{B(1)} \frac{B(z)}{P(z)} \tag{13.88}$$

13.2 Discrete Internal Model Control

Discrete internal model control (Garcia and Morari, 1982), (Morari and Zafiriou, 1989) is very similar with regard to its block diagram (Fig. 13.10) and the principle of continuous internal model control explained in Sect. 13.2. It presents very good robustness qualities, that is, faced with variations in structure or parameters of the process model. It is also possible to use it by taking into account constraints on the inputs and the states (Rotea and Marchetti, 1987). However, internal model control is, in principle, reserved for open-loop stable processes. In the case of open-loop unstable processes, it is necessary to first proceed to a classical stabilizing feedback.

The signals represented in the block diagram in Figure 13.10 are sampled signals. The process assumed to be open-loop stable, is represented by a discrete transfer function $G(z)$ and subjected to a disturbance $d(t)$. In practice, it would be the actual process. The process model is the discrete transfer function $\tilde{G}(z)$ while the internal model controller has the transfer function $G_c(z)$. A robustness filter with transfer function $F(z)$ is, in general, included in the feedback loop.

When the model is perfect, the feedback signal $\tilde{d}(t)$ in the loop is only the disturbance $d(t)$. Thus, internal model control appears as an open-loop system, implying the necessity of the stability of the process model G and the controller G_c. When the model and the process differ, the feedback signal $\tilde{d}(t)$ contains information concerning the model error, and it is possible to obtain some robustness by acting on $\tilde{d}(t)$ through the robustness filter $F(z)$.

According to the block diagram, including the filter $F(q)$, the following relations are obtained

$$\begin{aligned} u(t) &= \frac{1}{1 + G_c(q)\,(G(q) - \tilde{G}(q))}\,(G_c(q)\,y_r(t) - F(q)\,G_c(q)\,d(t)) \\ y(t) &= \frac{G(q)\,G_c(q)}{1 + F(q)\,G_c(q)\,(G(q) - \tilde{G}(q))}\,y_r(t) + \frac{1 - F(q)\,\tilde{G}(q)\,G_c(q)}{1 + F(q)\,G_c(q)\,(G(q) - \tilde{G}(q))}\,d(t) \end{aligned} \tag{13.89}$$

Suppose that $F(z) = 1$. From this equation, it results that a perfect control minimizing the sum of the error squares, either in regulation or set point tracking, would correspond to the following controller

$$G_c(z) = \frac{1}{\tilde{G}(z)} \tag{13.90}$$

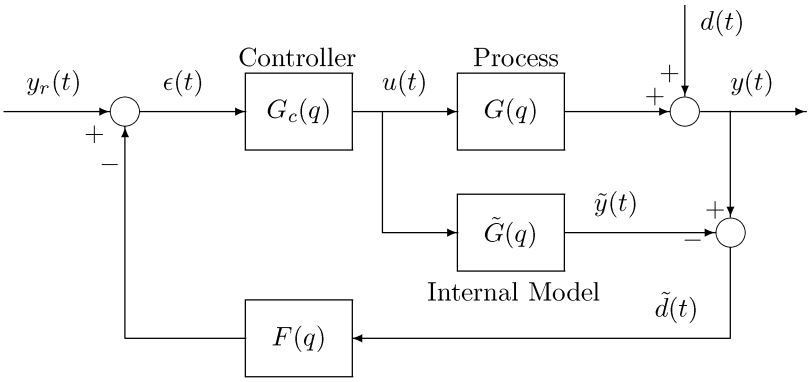


Figure 13.10: Discrete internal model control

Moreover, this controller would give zero deviation with respect to the set point. The same controller would also ensure perfect disturbance rejection. However, in general, such a controller is neither stable nor physically realizable.

The process is modelled by a transfer function expressed with respect to the poles and zeros

$$\tilde{G}(z) = z^{-d} \frac{\prod_{i=1}^{n-1}(z - z_i)}{\prod_{i=1}^{n}(z - p_i)} \tag{13.91}$$

where z_i represents a zero and p_i a pole of G.

The model $\tilde{G}(z)$ is factorized as

$$\tilde{G}(z) = \tilde{G}_+(z)\, \tilde{G}_-(z) \tag{13.92}$$

where $\tilde{G}_+(z)$ contains all the time delays (whose inversion would correspond to a prediction) and the zeros outside the unit circle (whose inversion would lead to an unstable controller), and thus includes all characteristics of nonminimum-phase type.

This factorization is not unique, but Garcia and Morari (1982) recommend choosing

$$\tilde{G}_+(z) = z^{-d-1} \prod_{i=1}^{n-1} \left(\frac{z - z_i}{z - \hat{z}_i} \right) \left(\frac{1 - \hat{z}_i}{1 - z_i} \right) \tag{13.93}$$

where d is the effective delay of the process and z_i represent the $n-1$ zeros of $\tilde{G}(z)$, while $\hat{z}_i$ are their images inside the unit circle

$$\hat{z}_i = \begin{cases} z_i & \text{if: } |z_i| \leq 1 \\ 1/z_i & \text{if: } |z_i| > 1 \end{cases} \tag{13.94}$$

Equation (13.93) thus retains in $\tilde{G}_+(z)$ only the unstable zeros and the time delays. The transfer function of the controller is then chosen as

$$G_c(z) = \frac{1}{\tilde{G}_-(z)} \tag{13.95}$$

In fact, the following rules (Zafiriou and Morari, 1986) will be used for the controller synthesis

- As $\tilde{G}(z)$ is stable, the n poles p_i of $\tilde{G}(z)$ become zeros of $G_c(z)$.
- The n_S stable and well-damped zeros z_i^S of $\tilde{G}(z)$ become poles of $G_c(z)$, while the n_I unstable zeros z_i^I with a positive real part of $\tilde{G}(z)$ are replaced by their inverse $1/z_i^I$, which will thus be stable poles of $G_c(z)$.
- The n^0 unstable zeros z_i^0 with a negative real part or badly damped zeros of $\tilde{G}(z)$ are not taken into account in $G_c(z)$.
- Poles at the origin are introduced in order to avoid anticipation.
- The gain of controller $G_c(z)$ is such that

$$G_c(1) = \frac{1}{\tilde{G}(1)} \tag{13.96}$$

By using these different rules, we obtain the following transfer function for the controller

$$G_c(z) = \frac{(-1)^{n^I}}{b_0 \prod_{i=1}^{n^0}(1 - z_i^0) \prod_{i=1}^{n^I} z_i^I} z^{-n+n^I+n^S} \frac{\prod_{i=1}^{n}(z - p_i)}{\prod_{i=1}^{n^S}(z - z_i^S) \prod_{i=1}^{n^I}(z_i^I z - 1)} \tag{13.97}$$

where b_0 is the coefficient of the monomial of the highest degree of the numerator of $\tilde{G}(z)$ assuming that the denominator is monic.

A robustness filter $F(z)$ is included in the feedback loop to take into account the imperfections of the model compared to the process through the estimated disturbance $\tilde{d}(t)$. In these conditions, Eq. (13.89) becomes

$$y(t) = \frac{G(q)}{\tilde{G}_-(q) + F(q)\,(G(q) - \tilde{G}(q))}\, y_r(t) + \frac{\tilde{G}_-(q)(1 - F(q)\,\tilde{G}^+(q))}{\tilde{G}_-(q) + F(q)\,(G(q) - \tilde{G}(q))}\, d(t) \tag{13.98}$$

If the model is perfect: $\tilde{G}(q) = G(q)$, Eq. (13.98) becomes

$$y(t) = \tilde{G}_+(q)\, y_r(t) + (1 - F(q)\,\tilde{G}^+(q))d(t) \tag{13.99}$$

Assuming that the gain of the robustness filter is equal to 1: $F(1) = 1$, the asymptotic convergence of the output y towards the set point y_r is guaranteed if $\tilde{G}_+(1) = 1$.

When the model is different from the process, the constant disturbances are asymptotically rejected if $F(1)\,\tilde{G}^+(1) = 1$. Assuming that $F(1) = 1$, this is equivalent to $G_c(1)^{-1} = \tilde{G}(1)$. Moreover, the output converges asymptotically towards the set point if $F(1) = 1$ and $\tilde{G}^+(1) = 1$.

The characteristic equation deduced from Eq. (13.98) allows us to discuss the stability of the system

$$\tilde{G}_-(z) + F(z)\,(G(z) - \tilde{G}(z)) = 0 \tag{13.100}$$

The filter $F(z)$ will be chosen such that this equation has its roots conveniently located in the unit circle. Thus, it is possible to use an exponential robustness

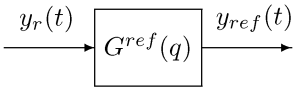


Figure 13.11: Reference model

filter which acts on the regulation dynamics

$$F(z) = \frac{z(1-\alpha)}{z-\alpha} \quad , \quad 0 \leq \alpha < 1 \tag{13.101}$$

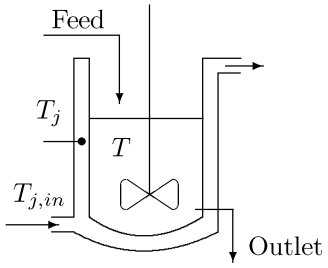
such that the increase of the filter time constant, thus the increase of parameter α, increases robustness.

In order to further improve the robustness, it is also recommended to filter the set point by a reference model $G^{ref}(q)$, of unit gain, which acts on the tracking dynamics (Fig. 13.11).

Thus, it is possible to separately set the regulation and tracking objectives.

The Smith predictor is a similar form of internal model control. Internal model control can also be considered from the point of view of linear quadratic control (Garcia and Morari, 1982). When the use of an exact, although stable, inverse in the controller brings excessive control variations, it may be interesting to replace the exact inverse by an approached inverse.

Example 13.3: **Discrete Internal Model Control of a Chemical Reactor**



The reactor discussed in Sect. 19.2 is used as the real process in this discrete internal model control, and constitutes a nonlinear simulation taking into account heat transfer and the chemical reaction when the latter occurs.

This reactor was previously identified by means of a very simple ARMAX model by the RELS method, with a sampling period $T_s = 5$ s, which gave

$$\tilde{G}(z) = \frac{1.916}{z-0.975}$$

The robustness filter corresponding to Eq. (13.101) has its constant $\alpha = 0.9$. The controller transfer function is equal to

$$G_c(z) = \frac{z-0.975}{1.916z} = \frac{0.522z-0.508}{z}$$

and corresponds to Eq. (13.97) by use of Eq. (13.93).

The reactor temperature set point is fixed at 300 K from 0 to 750 s, then 320 K from 750 s until the end. The reactor is not initially in steady state

so that the reactor temperature must in a first step reach the set point of 300 K. The control variable, corresponding to the opening percentage of a valve distributing the heat-conducting fluid between a hot heat exchanger and a cold heat exchanger, is bounded between 0 and 1. The output is affected by white noise of amplitude 0.5. The chemical reaction starts at 1500 s, and thus constitutes a disturbance for the system.

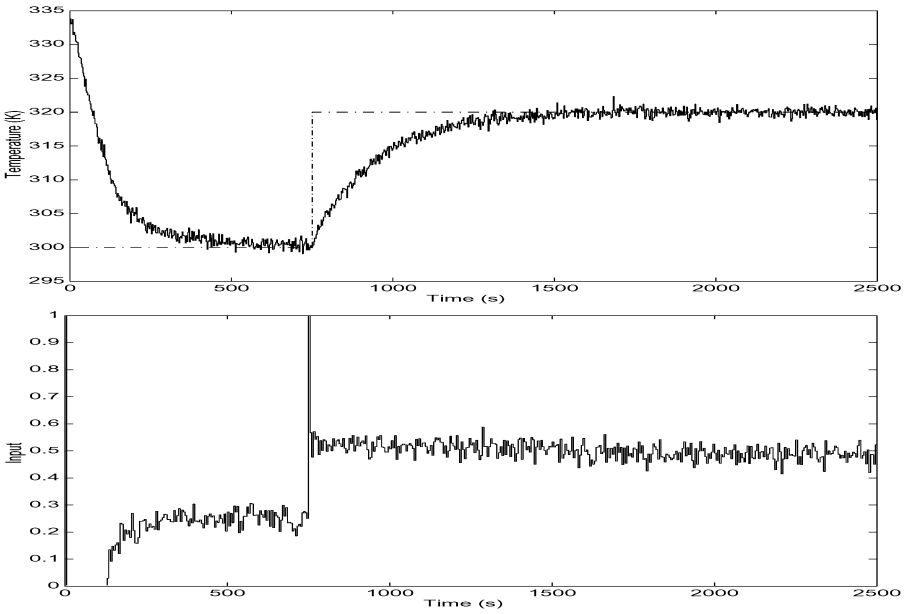


Figure 13.12: Internal model control of the chemical reactor with $\Delta H = -7 \times 10^4$ J/mol. Top: output temperature and set point. Bottom: control variable

The first simulation has been realized with a heat of reaction equal to $\Delta H = -7 \times 10^4$ J/mol. Figure 13.12 shows that the temperature very correctly reaches the first set point at 300 K, then the second set point at 320 K by presenting very little overshoot. The disturbance influence related to the chemical reaction is not visible. The control variable varies effectively between its lower and upper constraints. The control saturates when the physical demand of the heat-conducting system is at its highest, first in the cooling stage, then in the heating stage. In order to avoid too large variations of the control, the constant α of the robustness filter was voluntarily important, at the risk of losing a little performance. In these conditions, the control varies relatively slowly. The peaks at the start and the set point change could be avoided by introducing a reference trajectory.

In order to display firstly the difficulty of controlling exothermic reactors and secondly the robustness of internal model control, the heat of reaction was increased by a factor of 14 with respect to the concerned heat of reaction,

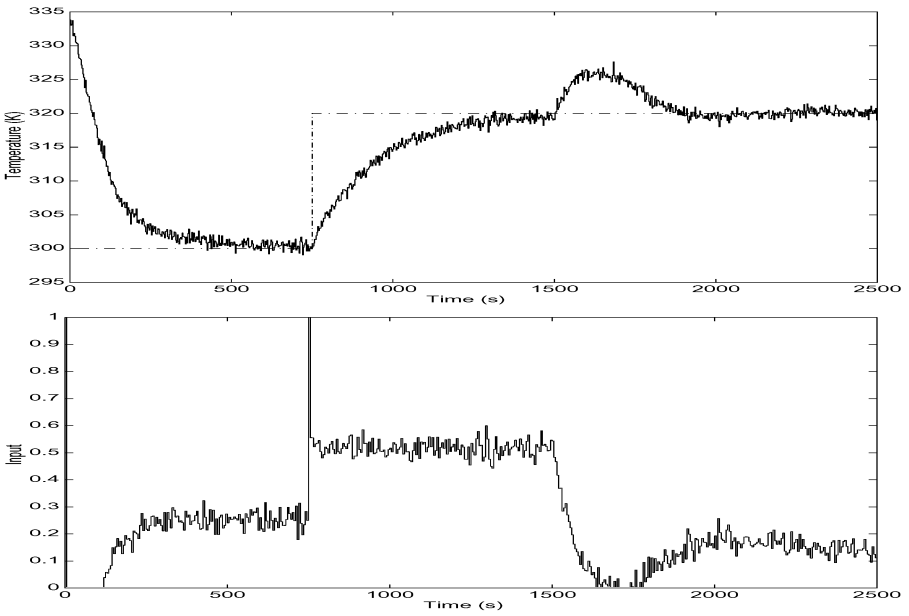


Figure 13.13: Internal model control of the chemical reactor at the limit of thermal runaway of the reactor in the case where the heat of reaction is large ($\Delta H = -10^6$ J/mol). Top: output temperature. Bottom: control variable

which is considerable. Figure 13.13 was thus obtained for $\Delta H = -10^6$ J/mol. In the first stage, after the beginning of the reaction (at 1500 s), a temperature deviation of the reactor with the set point occurs, then the system succeeds in controlling the temperature of the reactor contents. If the heat of reaction were increased to $\Delta H = -11 \times 10^5$ J/mol, an increasing deviation would appear because of the valve saturation and the temperature would be controlled no more. The physical capacity of cooling of the heat-conducting fluid would become insufficient and no control system could remedy that problem.

13.3 Generalities in Adaptive Control

(Goodwin and Sin, 1984) distinguish three types of control by increasing order of difficulty:

- Deterministic control, where disturbances are absent and the process model is known.
- Stochastic control, where disturbances are present and the process model is known.

– Adaptive control, where disturbances are present and the process model is completely specified.

In adaptive filtering, adaptive prediction or adaptive control, a problem of parametric estimation occurs. If the parameters are constant, the gain of the estimation algorithms converges towards zero, whereas if the parameters are time-varying, the algorithm must be able to follow the variations.

In general, two levels of adaptive control are distinguished:

- Indirect adaptive control: first, an estimation of the model parameters is realized, then the controller is calculated (Fig. 13.14). The control law is thus obtained in two separate stages.

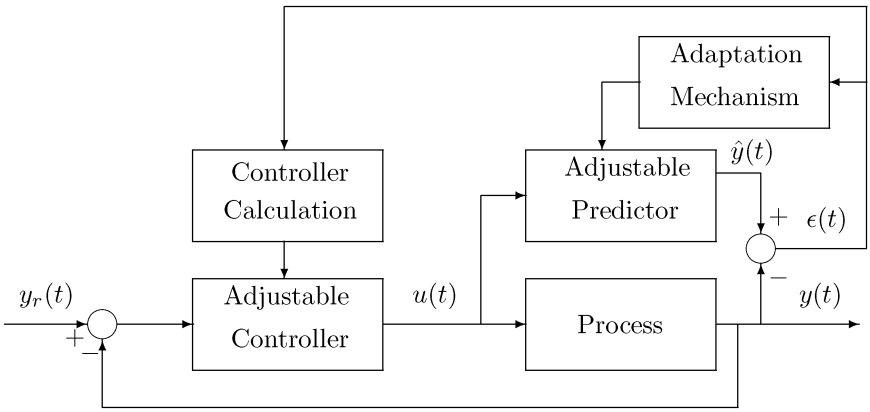


Figure 13.14: Indirect adaptive control

- Direct adaptive control: the right parameters of the controller are directly estimated (Fig. 13.15) by implicit integration of the model parameters.

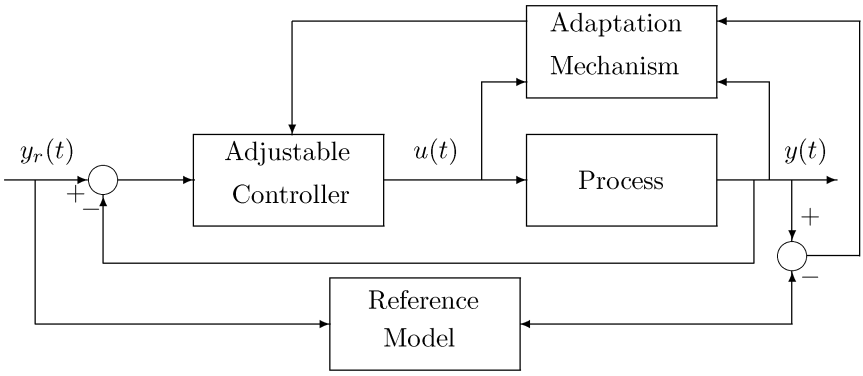


Figure 13.15: Direct adaptive control

The case where the controller parameters are already known in advance and preprogrammed (as in gain scheduling) is, in reality, regulation, not adaptive

control.

The problem of parametric identification was treated in Chap. 10. The adaptive control law is obtained by using the certainty equivalence principle, i.e. by replacing the model of the process by its admissible estimation when the control law is to be calculated.

Many control methods are susceptible to being used in indirect adaptive control: internal model control, pole-placement, pole and zero placement or model reference control.

Generalized predictive control (Sect. 15.4), with a reference model (Sect. 15.1), with a partial state reference model (Sect. 15.5) are direct adaptive control methods.

Adaptive control is very attractive because it is possible to hope to take into account slow variations of the process or disturbances affecting it. It imposes on us the taking of many precautions: convergence of the parameters of the process model, robustness of the on-line identification algorithm and stability of the control at each instant.

Many books are devoted to adaptive control and it is not possible to cite all of them: Aström and Wittenmark (1989); Bastin and Dochain (1990); Bitmead et al. (1990); Dugard and Landau (1990); Goodwin and Sin (1984); Landau (1988); Landau and Dugard (1986); Sastry and Bodson (1989); Watanabe (1992).

Some review papers are: Aström (1983); Aström and Wittenmark (1973); Aström et al. (1977); Garcia et al. (1989); Landau (1974); Seborg et al. (1986).

References

K.J. Aström. Robustness of a design method based on assignment of poles and zeros. *IEEE Trans. Automat. Control*, AC-25:588–591, 1980.

K.J. Aström. Theory and applications of adaptive control - A survey. *Automatica*, 19(5):471–486, 1983.

K.J. Aström and B. Wittenmark. On self tuning regulators. *Automatica*, pages 185–199, 1973.

K.J. Aström and B. Wittenmark. *Adaptive Control*. Addison-Wesley, New York, 1989.

K.J. Aström, U. Borisson, L. Ljung, and B. Wittenmark. Theory and applications of self-tuning regulators. *Automatica*, 13:457–476, 1977.

G. Bastin and D. Dochain. *On-Line Estimation and Adaptive Control of Bioreactors*. Elsevier, Amsterdam, 1990.

R. R. Bitmead, M. Gevers, and V. Wertz. *Adaptive Optimal Control, The Thinking Man's GPC*. Prentice Hall, New York, 1990.

P. De Larminat. *Automatique, Commande des Systèmes Linéaires*. Hermès, Paris, 1993.

L. Dugard and I.D. Landau, editors. *Commande Adaptative des Systèmes. Théorie, Méthodologie, Applications*, ENSIEG, BP 46, 38402 St-Martin d'Hères (France), 1990. Laboratoire d'Automatique de Grenoble.

C.E. Garcia and M. Morari. Internal model control. 1. A unifying review and some new results. *Ind. Eng. Chem. Process Des. Dev.*, 21:308–323, 1982.

C.E. Garcia, D.M. Prett, and M. Morari. Model predictive control: Theory and practice - a survey. *Automatica*, 25(3):335–348, 1989.

G.C. Goodwin and K.S. Sin. *Adaptive Filtering, Prediction and Control*. Prentice Hall, Englewood Cliffs, 1984.

R. Isermann. *Digital Control Systems*, volume II. Stochastic Control, Multivariable Control, Adaptive Control, Applications. Springer-Verlag, 2nd edition, 1991.

I.D. Landau. A survey of model reference adaptive techniques - Theory and applications. *Automatica*, 10:353–379, 1974.

I.D. Landau. *Identification et Commande des Systèmes*. Hermès, Paris, 1988.

I.D. Landau. *System Identification and Control Design*. Prentice Hall, Englewood Cliffs, 1990.

I.D. Landau and L. Dugard. *Commande Adaptative*. Masson, 1986.

I.D. Landau, R. Lozano, and M. M'Saad, editors. *Adaptive Control*. Springer-Verlag, London, 1997.

R.H. Middleton and G.C. Goodwin. *Digital Control and Estimation*. Prentice Hall, Englewood Cliffs, 1990.

M. Morari and E. Zafiriou. *Robust Process Control*. Prentice Hall, 1989.

K.S. Narendra and A.M. Annaswamy. *Stable Adaptive Systems*. Prentice Hall, Englewood Cliffs, 1989.

M.A. Rotea and J.L. Marchetti. Internal model control using the linear quadratic regulator theory. *Ind. Eng. Chem. Res.*, 26:577–581, 1987.

S. Sastry and M. Bodson. *Adaptive Control - Stability, Convergence, and Robustness*. Prentice Hall, New Jersey, 1989.

D.E. Seborg, T.F. Edgar, and S.L. Shah. Adaptive control strategies for process control: a survey. *AIChE J.*, 32(6):881–913, 1986.

M. Vidyasagar. *Control Systems Synthesis: A Factorization Approach*. MIT Press, Cambridge, Massachusetts, 1985.

K. Watanabe. *Adaptive Estimation and Control*. Prentice Hall, London, 1992.

E. Zafiriou and M. Morari. Design of robust digital controllers and sampling-time selection for SISO systems. *Int. J. Cont.*, 44(3):711–735, 1986.

Chapter 14

Optimal Control

14.1 Introduction

Frequently, the engineer in charge of a process is faced with optimization problems. In fact, this may cover relatively different ideas.

For example, consider a chemical reaction studied in a laboratory; the chemist seeks the best kinetic parameters: stoichiometries, reaction orders, rate constants, activation energies, i.e. the parameters that will help to represent the reacting system optimally. Then, the chemist realizes a static optimization, which most often consists of minimizing the distance between a set of experimental data and the corresponding prediction given by the model.

Now, consider the engineer in charge of a process, thus responsible for the real plant production. The laboratory chemist will have given to the engineer the information which seemed necessary concerning the reaction and has proposed a recipe for the experimental operation guide. The engineer knows that the reactor, being continuous, batch or fed-batch, and more or less similar to a perfectly stirred reactor or a plug-flow reactor, will behave, in reality, relatively differently to the lab reactor and will be closer to the pilot reactor, if the latter was previously operated. For example, he or she knows, that the reactive feed flow rate profile for a fed-batch reactor and the temperature or pressure profile to be followed for a batch reactor, will have an influence on the yield, the selectivity or the product quality. The engineer wishing to optimize production must then seek a time profile and realize a dynamic optimization with respect to the variables which he or she can manipulate, while respecting the constraints of the system such as the bounds on temperature and temperature rise rate, the constraints related to the possible runaway of the reactor. Similarly, an engineer realizing a reaction in a tubular reactor can seek the optimal temperature profile along the reactor. In the latter case, it is a spatial optimization very close to the dynamic optimization where the time is replaced by the abscissa along the reactor. The profile thus determined is calculated in open loop and will be applied as the set point in closed loop, which may lead to deviations between the effective result and the desired result. The di-

rect closed-loop calculation of the profile in the nonlinear case is not studied here; on the contrary, the linear case is treated in linear quadratic control and Gaussian linear quadratic control.

In a continuous process, problems of dynamic optimization can also be considered concerning the process changes from the nominal regime. For example, the quality of the raw petroleum feeding the refineries changes very often. The economic optimization realized off-line imposes set point variations on the distillation columns. An objective for the engineer can be to find the optimal profile to be followed during the change from one set of set points to another set.

In all cases, the engineer needs a dynamic model that is sufficiently representative of the behaviour of the process, and which is also of a reasonable complexity with respect to the difficulty of the mathematician and numerical task of solving.

Among the criteria that the engineer wishes to optimize, of course, can be found the reaction yield or the selectivity, and also the end-time taken to reach a given yield, or any technical-economic criterion which simultaneously takes into account technical objectives, production or investment costs.

Optimal control is the formulation of the dynamic optimization methods in the framework of a control problem.

14.2 Problem Statement

The optimal control problem is first set in continuous time. The studied system is assumed to be nonlinear. Applications which will follow in control will be developed only for linear systems.

The fixed aim in this problem is the determination of the control $\boldsymbol{u}(t)$ minimizing a criterion $J(\boldsymbol{u})$ while verifying initial and final conditions and respecting constraints. The optimal control thus denoted by $\boldsymbol{u}^*(t)$ makes the state $\boldsymbol{x}(t)$ follow a trajectory $\boldsymbol{x}^*(t)$ which must belong to the set of admissible trajectories.

The formulation of the optimal control problem is the following:
Consider a system described in state space by the set of differential equations

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t)) \qquad ; \quad t_0 \leq t \leq t_f \tag{14.1}$$

with $\boldsymbol{x}$ being a state vector of dimension n and $\boldsymbol{u}$ a control vector of dimension m. The system is subjected to initial and final conditions, called terminal (or at the boundaries)

$$\boldsymbol{k}(\boldsymbol{x}(t_0), t_0) = 0 \quad ; \; \boldsymbol{l}(\boldsymbol{x}(t_f), t_f) = 0 \tag{14.2}$$

Moreover, the system can be subjected to instantaneous inequality constraints

$$\boldsymbol{p}(\boldsymbol{x}(t), \boldsymbol{u}(t), t) \leq 0 \quad \forall t \tag{14.3}$$

or integral constraints (depending only on t_0 and t_f)

$$\int_{t_0}^{t_f} q(\boldsymbol{x}(t), \boldsymbol{u}(t), t) dt \leq 0 \tag{14.4}$$

The question is to find the set of the admissible controls $\boldsymbol{u}(t)$ which minimize a technical or economic performance criterion $J(\boldsymbol{u})$

$$J(\boldsymbol{u}) = G(\boldsymbol{x}(t_0), t_0, \boldsymbol{x}(t_f), t_f) + \int_{t_0}^{t_f} F(\boldsymbol{x}(t), \boldsymbol{u}(t), t)\, dt \tag{14.5}$$

G is called the algebraic part of the criterion. F is a functional[1]. Frequently, the initial instant is taken as $t_0 = 0$.

The performance index J depends on the type of problem. Some examples are:

- For a minimum-time problem (minimize t_f)

$$J = \int_0^{t_f} dt = t_f \tag{14.6}$$

- Minimization of the state error variance

$$J = \int_0^{t_f} (\boldsymbol{x}(t) - \boldsymbol{x}^{\text{ref}}(t))^2 dt \tag{14.7}$$

  where $\boldsymbol{x}^{\text{ref}}$ is the reference state.
- Combination of performance indices

$$J = \int_0^{t_f} dt + \mu \int_0^{t_f} (\boldsymbol{x}(t) - \boldsymbol{x}^{\text{ref}}(t))^2 dt \tag{14.8}$$

  where μ is a weighting factor.
- Weighted sum of the state error variance and the control error variance

$$J = \int_0^{t_f} \mu_1 \left[\boldsymbol{x}(t) - \boldsymbol{x}^{\text{ref}}(t)\right]^2 dt + \mu_2 \left[\boldsymbol{u}(t) - \boldsymbol{u}^{\text{ref}}(t)\right]^2 dt \tag{14.9}$$

- Use of a nonlinear functional

$$J = \int_0^{t_f} F(\boldsymbol{x}(t), \boldsymbol{u}(t), t)\, dt \tag{14.10}$$

- Combination with a functional

$$J = G(\boldsymbol{x}(0), \boldsymbol{u}(0), \boldsymbol{x}(t_f), \boldsymbol{u}(t_f)) + \int_0^{t_f} F(\boldsymbol{x}(t), \boldsymbol{u}(t), t)\, dt \tag{14.11}$$

Notice that an inequality constraint such as (14.3) can be replaced (Boudarel et al., 1969) by an equality constraint, by adding an auxiliary function according to Valentine's method, as

$$\boldsymbol{p}(\boldsymbol{x}(t), \boldsymbol{u}(t), t) + \boldsymbol{y}^2(t) = 0 \quad \forall t \tag{14.12}$$

[1] A functional is a function of functions: the function $F(\boldsymbol{x}(t), \boldsymbol{u}(t), t)$ depends on functions $\boldsymbol{x}(t)$ and $\boldsymbol{u}(t)$.

Similarly, the inequality (14.4) becomes

$$\int_{t_0}^{t_f} [q(\boldsymbol{x}(t), \boldsymbol{u}(t), t) + z(t)^2] dt = 0 \tag{14.13}$$

In fact, by introducing the variable $w(t)$ such that

$$w(t) = \int_{t_0}^{t} [q(\boldsymbol{x}(\tau), \boldsymbol{u}(\tau), \tau) + z(\tau)^2] d\tau \tag{14.14}$$

the integral constraint is transformed into an instantaneous constraint

$$\dot{w}(t) - [q(\boldsymbol{x}(t), \boldsymbol{u}(t), t) + z(t)^2] = 0 \tag{14.15}$$

where w must verify the terminal conditions: $w(t_0) = 0$, $w(t_f) = 0$.

In this very general form, this problem makes use of the equality constraints corresponding to the state differential equations, the terminal equality constraints, possibly instantaneous or integral inequality constraints, and m independent functions, which are the controls $\boldsymbol{u}(t)$. The term $G(\boldsymbol{x}(t_0), t_0, \boldsymbol{x}(t_f), t_f)$ represents a contribution of the terminal conditions to the criterion whereas the integral term of Eq. (14.5) represents a time-accumulation contribution.

Several methods allow us to solve this type of problem: variational methods (Kirk, 1970), Pontryagin maximum principle (Pontryaguine et al., 1974), Bellman dynamic programming (Bellman, 1957). The books cited here (Borne et al., 1990; Boudarel et al., 1969; Bryson, 1999; Bryson and Ho, 1975; Feldbaum, 1973; Pun, 1972; Ray and Szekely, 1973) propose compared approaches.

Variational methods are first presented in a mathematical form which can be qualified as dynamic optimization, in order to distinguish the tools independently from the control problem. Then, the control problem is treated by a specification of some variables.

14.3 Variational Method in the Mathematical Framework

In the most general mathematical form, in the classical variational method, the performance index to be minimized with respect to the vector of functions $\boldsymbol{f}$ (of dimension n) of the variable x is the following

$$J(\boldsymbol{f}) = G(x_0, \boldsymbol{f}(x_0), x_1, \boldsymbol{f}(x_1)) + \int_{x_0}^{x_1} F(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x)) dx \tag{14.16}$$

where $\boldsymbol{f}$ and $\dot{\boldsymbol{f}}$ are considered as independent variables. The vector of functions $\boldsymbol{f}$ must satisfy the following equations

$$\phi_i(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x)) = 0 \qquad i = 1, \ldots, n_\phi < n \tag{14.17}$$

$$k_j(x_0, \boldsymbol{f}(x_0)) = 0 \qquad j = 1, \ldots, n_0 \tag{14.18}$$

$$l_j(x_1, \boldsymbol{f}(x_1)) = 0 \qquad j = n_0 + 1, \ldots, n_0 + n_1 \leq 2n + 2 \tag{14.19}$$

In this form, the problem is called a Bolza problem. If the functional F of Eq. (14.16) is zero, this is a Mayer problem. If G is zero, it is a Lagrange problem. General comments:

- Notice that in the performance index J, the term G depends only on the initial (or lower) limit x_0 and on the final (or upper) limit x_1, while the integral term depends on the whole history between the initial and final limits.
- Equation (14.17) is a system of differential equations.
- Equation (14.18) represents the constraints at the initial limit.
- Equation (14.19) represents the constraints at the final limit.
- If we define an additional variable

$$f_{n+1}(x) = G(x_0, \boldsymbol{f}(x_0), x, \boldsymbol{f}(x)) + \int_{x_0}^{x} F(\xi, \boldsymbol{f}(\xi), \dot{\boldsymbol{f}}(\xi))d\xi \tag{14.20}$$

equation equivalent to

$$\dot{f}_{n+1}(x) = G_{\boldsymbol{f}}^T \dot{\boldsymbol{f}} + G_x + F(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x)) \qquad \text{with: } f_{n+1}(x_0) = 0 \tag{14.21}$$

and the performance index J equal to

$$J = f_{n+1}(x_1) = G(x_0, \boldsymbol{f}(x_0), x_1, \boldsymbol{f}(x_1)) + \int_{x_0}^{x_1} F(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x))dx \tag{14.22}$$

the problem is analogous to the Pontryagin method.

14.3.1 Variation of the Criterion

Assuming that $\boldsymbol{f}^*$ is the optimal solution belonging to the domain of admissible trajectories, we study the influence of the variations[2] of $\boldsymbol{f}$ in the neighbourhood of $\boldsymbol{f}^*$ on the criterion J.

[2] Several mathematical relations are useful:
a) We denote by y_z the partial derivative $\partial y/\partial z$, where z is a scalar. If y is scalar and $\boldsymbol{z}$ a vector, the notation $y_{\boldsymbol{z}}$ is the gradient vector of the partial derivatives $\partial y/\partial z_i$. If $\boldsymbol{y}$ and $\boldsymbol{z}$ are vectors, the notation $\boldsymbol{y}_{\boldsymbol{z}}$ represents the Jacobian matrix of the current element $\partial y_i/\partial z_j$.
b) The derivative with respect to $\boldsymbol{f}$ of the integral with fixed boundaries

$$I = \int_{x_0}^{x_1} F(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x))dx$$

is equal to

$$\frac{dI}{d\boldsymbol{f}} = \int_{x_0}^{x_1} \left[F_{\boldsymbol{f}} - \frac{d}{dx} F_{\dot{\boldsymbol{f}}} \right] dx$$

c) According to the Euler-Lagrange lemma (Cartan, 1967), if $\boldsymbol{C}(x)$ is a continuous function (vector) on $[a, b]$ verifying

$$\int_a^b \boldsymbol{C}^T(x)\boldsymbol{v}(x)dx = 0$$

for all function (vector) $\boldsymbol{v}(x)$ which is continuous and becomes zero at the boundaries, then $\boldsymbol{C}(x)$ is zero everywhere on $[a, b]$.

In the most general case, supposing that the boundaries x_0 and x_1 are not fixed, noting that $f_0 = f(x_0)$ and likewise for f_1, and $F_0 = F(x_0, \boldsymbol{f}(x_0), \dot{\boldsymbol{f}}(x_0))$ and similarly for F_1, the variation of criterion J of Eq. (14.16) related to a variation of these boundaries is equal to

$$\begin{aligned}\delta J &= \int_{x_0}^{x_1} \left[\left(\frac{\partial F}{\partial \boldsymbol{f}}\right)^T \delta \boldsymbol{f} + \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)^T \delta \dot{\boldsymbol{f}} \right] dx + [F_1\, \delta x_1 - F_0\, \delta x_0] \\ &+ \left[\left(\frac{\partial G}{\partial x_0}\right) \delta x_0 + \left(\frac{\partial G}{\partial \boldsymbol{f}_0}\right)^T \delta \boldsymbol{f}_0 \right] + \left[\left(\frac{\partial G}{\partial x_1}\right) \delta x_1 + \left(\frac{\partial G}{\partial \boldsymbol{f}_1}\right)^T \delta \boldsymbol{f}_1 \right]\end{aligned} \tag{14.23}$$

The second part of the integral term can be expressed as

$$\begin{aligned}\int_{x_0}^{x_1} \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)^T \delta \dot{\boldsymbol{f}}\, dx &= \int_{x_0}^{x_1} \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)^T \frac{d}{dx}[\delta \boldsymbol{f}] dx \\ &= \left[\left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)^T \delta \boldsymbol{f} \right]_{x_0}^{x_1} - \int_{x_0}^{x_1} \frac{d}{dx} \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)^T \delta \boldsymbol{f}\, dx \\ &= \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)_1^T \delta \boldsymbol{f}(x_1) - \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)_0^T \delta \boldsymbol{f}(x_0) - \int_{x_0}^{x_1} \frac{d}{dx} \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)^T \delta \boldsymbol{f}\, dx\end{aligned} \tag{14.24}$$

On the other hand, the following relations expressing the variation of the terminal functions as the sum of two contributions are used

$$\delta \boldsymbol{f}_0 = \delta \boldsymbol{f}(x_0) + \dot{\boldsymbol{f}}(x_0)\delta(x_0) \quad \text{and} \quad \delta \boldsymbol{f}_1 = \delta \boldsymbol{f}(x_1) + \dot{\boldsymbol{f}}(x_1)\delta(x_1) \tag{14.25}$$

Using an integration by parts of the second term of the integral, the criterion variation equation can be transformed into

$$\begin{aligned}\delta J &= \int_{x_0}^{x_1} \left[\frac{\partial F}{\partial \boldsymbol{f}} - \frac{d}{dx} \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right) \right]^T \delta \boldsymbol{f} dx \\ &+ \left[\frac{\partial G}{\partial x_1} + \left(F - \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)^T \dot{\boldsymbol{f}} \right)_1 \right] \delta x_1 + \left[\frac{\partial G}{\partial \boldsymbol{f_1}} + \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)_1 \right]^T \delta \boldsymbol{f}_1 \\ &- \left[-\frac{\partial G}{\partial x_0} + \left(F - \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)^T \dot{\boldsymbol{f}} \right)_0 \right] \delta x_0 - \left[-\frac{\partial G}{\partial \boldsymbol{f_0}} + \left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)_0 \right]^T \delta \boldsymbol{f}_0\end{aligned} \tag{14.26}$$

Moreover, the variations at the boundaries are dependent because of the constraints (14.18) and (14.19), giving the relations

$$\begin{aligned}\left(\frac{\partial \boldsymbol{k}}{\partial x}\right)_0 \delta x_0 + \left(\frac{\partial \boldsymbol{k}}{\partial \boldsymbol{f}}\right)_0 \delta \boldsymbol{f}_0 = 0 \\ \left(\frac{\partial \boldsymbol{l}}{\partial x}\right)_1 \delta x_1 + \left(\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{f}}\right)_1 \delta \boldsymbol{f}_1 = 0\end{aligned} \tag{14.27}$$

A necessary, but not sufficient, condition of the minimum of the criterion is

$$\delta J = 0 \qquad , \quad \forall\, \delta \boldsymbol{f},\ \delta \boldsymbol{f}_0,\ \delta \boldsymbol{f}_1,\ \delta x_0,\ \delta x_1 \tag{14.28}$$

14.3.2 Variational Problem without Constraints, Fixed Boundaries

Consider the simple case where the criterion to be optimized is simply in the form

$$J = \int_{x_0}^{x_1} F(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x))dx \tag{14.29}$$

and the boundaries x_0 and x_1 are fixed. We seek the admissible optimal trajectory $\boldsymbol{f}^*$, minimizing the criterion J with respect to $\boldsymbol{f}$

$$\boldsymbol{f}^* = \arg\left\{\min_{\boldsymbol{f}} J\right\} \tag{14.30}$$

From Eq. (14.26), using the Euler-Lagrange lemma, we deduce the necessary Euler conditions so that J has a local extremum in $\boldsymbol{f}^*$ (necessary condition of stationarity), i.e. the following vector is zero

$$\left(\frac{\partial F}{\partial \boldsymbol{f}}\right)_* - \left(\frac{d}{dx}\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)_* = 0 \tag{14.31}$$

The system (14.31) can be written again in the following form

$$\frac{\partial F}{\partial \boldsymbol{f}} - \frac{\partial\left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)}{\partial x} - \frac{\partial\left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)}{\partial f}\dot{\boldsymbol{f}} - \frac{\partial\left(\frac{\partial F}{\partial \dot{\boldsymbol{f}}}\right)}{\partial \dot{\boldsymbol{f}}}\ddot{\boldsymbol{f}} = 0 \tag{14.32}$$

and thus constitutes n second-order differential equations. The $2n$ degrees of freedom are filled by the n initial and n final conditions.

14.3.3 Variational Problem with Constraints, General Case

Consider the simple case where the criterion to be optimized is

$$J = G(x_0, \boldsymbol{f}(x_0), x_1, \boldsymbol{f}(x_1)) + \int_{x_0}^{x_1} F(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x))dx \tag{14.33}$$

and the boundaries x_0 and x_1 are not fixed. We seek the admissible optimal trajectory $\boldsymbol{f}^*$, minimizing the criterion J with respect to $\boldsymbol{f}$

$$\boldsymbol{f}^* = \arg\left\{\min_{\boldsymbol{f}} J\right\} \tag{14.34}$$

When the problem includes equality constraints such as Eq. (14.17)

$$\boldsymbol{\phi}(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x)) = 0 \quad \forall x \in [x_0, x_1] \tag{14.35}$$

or inequality constraints transformed into equality constraints such as Eq. (14.12)

$$\boldsymbol{p}(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x)) + \boldsymbol{y}^2(x) = 0 \quad \forall x \in [x_0, x_1] \tag{14.36}$$

and such as Eq. (14.13)

$$\int_{x_0}^{x_1} [q(x, \boldsymbol{f}(x), \dot{\boldsymbol{f}}(x)) + z(x)^2] dx = 0 \tag{14.37}$$

we apply the *Euler conditions* (as well as all the terminal conditions, discontinuity conditions and conditions relative to the second variations) to the augmented function F^a

$$\begin{aligned} F^a(x, \boldsymbol{f}, \dot{\boldsymbol{f}}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}) = \quad & F(x, \boldsymbol{f}, \dot{\boldsymbol{f}}) \\ & + \boldsymbol{\lambda}^T(x)\, \boldsymbol{\phi} + \boldsymbol{\mu}^T(x)\, [\boldsymbol{p} + \boldsymbol{y}^2] + \boldsymbol{\nu}^T(x)\, [\dot{\boldsymbol{w}} - \boldsymbol{q} - \boldsymbol{z}^2] \end{aligned} \tag{14.38}$$

obtained by introducing for each constraint a Lagrange or Kuhn-Tucker parameter. In this case, the variables concerned by the Euler conditions are: $\boldsymbol{f}$, $\boldsymbol{y}$, $\boldsymbol{z}$, $\boldsymbol{w}$. The Euler conditions applied to the augmented function F^a give the following equations

– With respect to variable $\boldsymbol{f}$

$$\begin{aligned} & \left(\frac{\partial F^a}{\partial \boldsymbol{f}}\right)_* - \left(\frac{d}{dx}\frac{\partial F^a}{\partial \dot{\boldsymbol{f}}}\right)_* = 0 \quad \Longrightarrow \\ & F_{\boldsymbol{f}} + \boldsymbol{\phi}_{\boldsymbol{f}}^T \boldsymbol{\lambda} + \boldsymbol{p}_{\boldsymbol{f}}^T \boldsymbol{\mu} - \boldsymbol{q}_{\boldsymbol{f}}^T \boldsymbol{\nu} - \frac{d}{dx}(F_{\dot{\boldsymbol{f}}} + \boldsymbol{\phi}_{\dot{\boldsymbol{f}}}^T \boldsymbol{\lambda} + \boldsymbol{p}_{\dot{\boldsymbol{f}}}^T \boldsymbol{\mu} - \boldsymbol{q}_{\dot{\boldsymbol{f}}}^T \boldsymbol{\nu}) = 0 \end{aligned} \tag{14.39}$$

– With respect to variable $\boldsymbol{y}$

$$2\boldsymbol{\mu}^T \boldsymbol{y} = 0 \tag{14.40}$$

– With respect to variable $\boldsymbol{z}$

$$2\boldsymbol{\nu}^T \boldsymbol{z} = 0 \tag{14.41}$$

– With respect to variable $\boldsymbol{w}$

$$\dot{\boldsymbol{\nu}} = 0 \quad \Longrightarrow \boldsymbol{\nu} = \text{constant} \tag{14.42}$$

After applying the Euler conditions, the boundary conditions and the trajectory discontinuities must be taken into account.

Terminal Conditions

The constraints (14.18) acting on the initial boundary x_0 and (14.19) acting on the final boundary x_1 are called the terminal conditions. They express that the terminal values of functions $\boldsymbol{f}$, x, belong to hypersurfaces or, more simply, that relations link the terminal values of functions $\boldsymbol{f}(x)$.

Transversality conditions

Taking into account the variation of criterion J described by Eq. (14.26), the Euler condition of stationarity and the constraints acting on the boundary variations $\delta \boldsymbol{f}_0$, δx_0, $\delta \boldsymbol{f}_1$, δx_1, we obtain the transversality conditions, which are written as:

At the initial boundary x_0

$$\left[-\frac{\partial G}{\partial x_0}+\left(F^a-\left(\frac{\partial F^a}{\partial \dot{\boldsymbol{f}}}\right)^T\dot{\boldsymbol{f}}\right)_0\right]\delta x_0+\left[-\frac{\partial G}{\partial \boldsymbol{f}_0}+\left(\frac{\partial F^a}{\partial \dot{\boldsymbol{f}}}\right)_0\right]^T\delta \boldsymbol{f}_0=0$$
$$\text{with:}\quad \left(\frac{\partial \boldsymbol{k}}{\partial x}\right)_0\delta x_0+\left(\frac{\partial \boldsymbol{k}}{\partial \boldsymbol{f}}\right)_0\delta \boldsymbol{f}_0=0 \tag{14.43}$$

At the final boundary x_1

$$\left[\frac{\partial G}{\partial x_1}+\left(F^a-\left(\frac{\partial F^a}{\partial \dot{\boldsymbol{f}}}\right)^T\dot{\boldsymbol{f}}\right)_1\right]\delta x_1+\left[\frac{\partial G}{\partial \boldsymbol{f}_1}+\left(\frac{\partial F^a}{\partial \dot{\boldsymbol{f}}}\right)_1\right]^T\delta \boldsymbol{f}_1=0$$
$$\text{with:}\quad \left(\frac{\partial \boldsymbol{l}}{\partial x}\right)_1\delta x_1+\left(\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{f}}\right)_1\delta \boldsymbol{f}_1=0 \tag{14.44}$$

This means that the extremizing trajectory must have in the phase space of variables f_i the same slope as the trajectories $\boldsymbol{k}$ and $\boldsymbol{l}$ at the initial and final boundaries x_0 and x_1 respectively.

If the initial boundary x_0 is fixed, this leads to $\delta x_0=0$ (similarly, $\delta x_1=0$ if the final boundary x_1 is fixed). If the initial extremity of the trajectory $\boldsymbol{f}_0$ is fixed, this leads to $\delta \boldsymbol{f}_0=0$ (similarly, $\delta \boldsymbol{f}_1=0$ if the final extremity $\boldsymbol{f}_1$ is fixed).

Discontinuity Condition

This condition is also called the Weierstrass-Erdmann condition. The discontinuous extremizing trajectories are composed of sub-arcs joined by discontinuities. At these points, the partial derivatives of two sub-arcs must be equal

$$\left(\frac{\partial F^a}{\partial \dot{f}_i}\right)_-=\left(\frac{\partial F^a}{\partial \dot{f}_i}\right)_+\quad;\quad i=1,\ldots,n \tag{14.45}$$

$$\left(-F^a+\sum_{i=1}^n\frac{\partial F^a}{\partial \dot{f}_i}\dot{f}_i\right)_-=\left(-F^a+\sum_{i=1}^n\frac{\partial F^a}{\partial \dot{f}_i}\dot{f}_i\right)_+\quad;\quad i=1,\ldots,n \tag{14.46}$$

In optimal control, an extremizing trajectory thus can be composed of a first arc, where the control is saturated at its minimum value u_-, followed by an optimal arc and then followed by an arc where the control is saturated at its maximum value u_+. In practice, it is common that the control takes only the minimum and maximum values (bang-bang control).

Partial conclusion and first solution

After having applied the Euler equations and the terminal and discontinuity conditions, we have, in fact, realized a set of necessary conditions for obtaining a stationary solution, but not sufficient for the minimum solution that will be called a first solution.

Conditions relative to the second variations

Weierstrass-Erdmann and Legendre-Clebsch conditions are necessary conditions for the criterion J to be minimum. Weierstrass conditions are relative to large variations and Legendre conditions to small variations.

Weierstrass-Erdmann Condition
The necessary condition for the performance index J to be minimum is that the Weierstrass function W verifies

$$W = F^a(\boldsymbol{f}^*, \dot{\boldsymbol{f}}, x) - F^a(\boldsymbol{f}^*, \dot{\boldsymbol{f}}^*, x) - (\dot{\boldsymbol{f}} - \dot{\boldsymbol{f}}^*)^T \left(\frac{\partial F^a}{\partial \dot{\boldsymbol{f}}} \right)_* \geq 0 \tag{14.47}$$

at any point of the extremum arc, for all large variations $\Delta \dot{f}_i$ that are compatible with constraints ϕ_i.
Legendre-Clebsch Condition
This condition is relative to small variations $\delta \dot{f}$ in the neighbourhood of $\dot{f}^*$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 F^a}{\partial \dot{f}_i \partial \dot{f}_j} \, \delta \dot{f}_i \, \delta \dot{f}_j \geq 0 \tag{14.48}$$

If a maximum performance index were sought, the two previous inequalities would have an opposite sign.

By applying the conditions relative to the second variations, the optimal solution is completely determined.

14.3.4 Hamilton-Jacobi Equation

The problem formulation based on the use of the Hamiltonian allows us to express Euler equations in the canonical form of Hamilton equations. The Hamilton-Jacobi equation bearing on the criterion optimization leads to us obtaining the optimal trajectory. The discrete Hamilton-Jacobi equation can be very well compared to the Bellman optimality principle in dynamic programming (Sect. 14.5).

If, on the optimal trajectory, we introduce the variable $\boldsymbol{\psi}(x)$ defined as

$$\boldsymbol{\psi}(x) = \frac{\partial F}{\partial \dot{\boldsymbol{f}}} \tag{14.49}$$

the Euler condition (14.31) becomes

$$\frac{d}{dx} \boldsymbol{\psi}(x) = \frac{\partial F}{\partial \boldsymbol{f}} \tag{14.50}$$

Provided that the matrix

$$\frac{\partial^2 F}{\partial \dot{\boldsymbol{f}}^2} \tag{14.51}$$

is nonsingular, the function $\dot{\boldsymbol{f}}$ can be expressed as a solution of implicit Eq. (14.49) as

$$\dot{\boldsymbol{f}}(x) = \boldsymbol{p}(\boldsymbol{f}(x), \boldsymbol{\psi}(x), x) \tag{14.52}$$

We define the Hamiltonian function or Hamiltonian (similar to the Lagrangian function for classical optimization problems) as

$$H(\boldsymbol{f}(x), \boldsymbol{\psi}(x), x) = -F(\boldsymbol{f}(x), \dot{\boldsymbol{f}}(x), x) + \boldsymbol{\psi}^T \boldsymbol{p}(\boldsymbol{f}(x), \boldsymbol{\psi}(x), x) \tag{14.53}$$

The partial derivatives of the Hamiltonian are equal to

$$\begin{aligned} H_{\boldsymbol{f}} &= -F_{\boldsymbol{f}} - \dot{\boldsymbol{f}}_{\boldsymbol{f}}^T F_{\dot{\boldsymbol{f}}} + \boldsymbol{p}_{\boldsymbol{f}}^T \boldsymbol{\psi} = -F_{\boldsymbol{f}} - \boldsymbol{p}_{\boldsymbol{f}}^T F\boldsymbol{p} + \boldsymbol{p}_{\boldsymbol{f}}^T \boldsymbol{\psi} = -F_{\boldsymbol{f}} \\ H_{\boldsymbol{\psi}} &= -\dot{\boldsymbol{f}}_{\boldsymbol{\psi}}^T F_{\dot{\boldsymbol{f}}} + \boldsymbol{p} + \boldsymbol{p}_{\boldsymbol{\psi}}^T \boldsymbol{\psi} = -\boldsymbol{p}_{\boldsymbol{\psi}}^T F\boldsymbol{p} + \boldsymbol{p} + \boldsymbol{p}_{\boldsymbol{\psi}}^T \boldsymbol{\psi} = \boldsymbol{p} \\ H_x &= -F_x - F_{\boldsymbol{f}}^T \boldsymbol{f}_x - F_{\dot{\boldsymbol{f}}}^T \dot{\boldsymbol{f}}_x + \boldsymbol{\psi}_x^T \boldsymbol{p} + \boldsymbol{\psi}^T \boldsymbol{p}_x \\ &= -F_x - \dot{\boldsymbol{\psi}}^T \boldsymbol{p} - \boldsymbol{\psi}^T \boldsymbol{p}_x + \dot{\boldsymbol{\psi}}^T \boldsymbol{p} + \boldsymbol{\psi}^T \boldsymbol{p}_x = -F_x \end{aligned} \tag{14.54}$$

On the other side, using the first two expressions of the partial derivatives, the derivative of the Hamiltonian gives

$$\frac{dH}{dx} = H_{\boldsymbol{f}}^T \dot{\boldsymbol{f}} + H_{\boldsymbol{\psi}}^T \dot{\boldsymbol{\psi}} + H_x = -F_x \tag{14.55}$$

from which we get the Hamilton canonical conditions for optimality

$$\begin{aligned} \dot{\boldsymbol{f}} &= H_{\boldsymbol{\psi}} \\ \dot{\boldsymbol{\psi}} &= -H_{\boldsymbol{f}} \\ \frac{dH}{dx} &= -F_x \end{aligned} \tag{14.56}$$

The first two equations of this system constitute a set of $2n$ equations equivalent to Euler conditions.

In the neighbourhood of the optimal trajectory $\boldsymbol{f}^*$, realize a variation of the trajectory at the boundary x but with always verifying the final condition, and consider the criterion at the optimal trajectory $\boldsymbol{f}^*$

$$\mathcal{J}(\boldsymbol{f}^*, x) = G(\boldsymbol{f}_1^*, x_1) + \int_x^{x_1} F(\xi, \boldsymbol{f}^*(\xi), \dot{\boldsymbol{f}}^*(\xi)) d\xi \tag{14.57}$$

According to Eq. (14.26), the variation of criterion $\mathcal{J}$ in the neighbourhood of the optimal solution $\boldsymbol{f}^*$ is given by

$$\delta\mathcal{J} = -\left(F - F_{\dot{\boldsymbol{f}}}^T \dot{\boldsymbol{f}}^*\right) \delta x - F_{\dot{\boldsymbol{f}}}^T \delta \boldsymbol{f} = H(\boldsymbol{f}^*(x), \boldsymbol{\psi}(x), x)\delta x - \boldsymbol{\psi}^T(x)\, \delta \boldsymbol{f}(x) \tag{14.58}$$

hence

$$\begin{aligned} \mathcal{J}_x &= H(\boldsymbol{f}^*(x), \boldsymbol{\psi}(x), x) \\ \mathcal{J}_{\boldsymbol{f}} &= -\boldsymbol{\psi}(x) \end{aligned} \tag{14.59}$$

The system of Eqs. (14.59) provides the Hamilton-Jacobi equation

$$\mathcal{J}_x = H(\boldsymbol{f}^*(x), -\mathcal{J}_{\boldsymbol{f}}, x) \tag{14.60}$$

which is a first-order partial derivative equation which admits as a solution the integral $\mathcal{J}(\boldsymbol{f}^*, x)$ defined by Eq. (14.57). Along an optimal trajectory, the criterion solution of Eq. (14.60) is optimal. The boundary condition of $\mathcal{J}$ takes into account the end part which is eventually present in criterion (14.57)

$$\mathcal{J}(\boldsymbol{f}^*, x_1) = G(\boldsymbol{f}_1^*, x_1) \tag{14.61}$$

The transversality conditions are deduced from Eqs. (14.43) and (14.44): at the initial boundary x_0, where the constraint is satisfied for all δx_0 and $\delta \boldsymbol{f}_0$

$$\left(\frac{\partial \boldsymbol{k}}{\partial x}\right)_0 \delta x_0 + \left(\frac{\partial \boldsymbol{k}}{\partial \boldsymbol{f}}\right)_0 \delta \boldsymbol{f}_0 = 0 \tag{14.62}$$

the transversality condition is imposed

$$\left[-\frac{\partial G}{\partial x_0} - H_0\right] \delta x_0 + \left[-\frac{\partial G}{\partial \boldsymbol{f_0}} + \boldsymbol{\psi}_0\right]^T \delta \boldsymbol{f}_0 = 0 \tag{14.63}$$

at the final boundary x_1, where the constraint is satisfied for all δx_1 and $\delta \boldsymbol{f}_1$

$$\left(\frac{\partial \boldsymbol{l}}{\partial x}\right)_1 \delta x_1 + \left(\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{f}}\right)_1 \delta \boldsymbol{f}_1 = 0 \tag{14.64}$$

the transversality condition is imposed

$$\left[\frac{\partial G}{\partial x_1} - H_1\right] \delta x_1 + \left[\frac{\partial G}{\partial \boldsymbol{f_1}} + \boldsymbol{\psi}_1\right]^T \delta \boldsymbol{f}_1 = 0 \tag{14.65}$$

If x_1 is fixed, from Eq. (14.65), the frequently encountered condition results

$$\boldsymbol{\psi}(x_1) = -\frac{\partial G}{\partial \boldsymbol{f_1}} \tag{14.66}$$

14.4 Optimal Control

14.4.1 Variational Methods

As, henceforth, we place ourselves in the framework of optimal control, x plays the role of time and the variables f_k are divided in two types: state variables x_i $(1 \leq i \leq n)$ and control variables u_j $(1 \leq j \leq m)$, so that the new problem is formulated as:
Given a criterion

$$J(\boldsymbol{u}) = G(\boldsymbol{x}(t_0), \boldsymbol{u}(t_0), \boldsymbol{x}(t_f), \boldsymbol{u}(t_f)) + \int_{t_0}^{t_f} F(\boldsymbol{x}(t), \boldsymbol{u}(t), t)dt \tag{14.67}$$

we seek the optimal control trajectory $\boldsymbol{u}^*(t)$ that minimizes $J(\boldsymbol{u})$

$$\boldsymbol{u}^*(t) = \arg\left\{\min_{\boldsymbol{u}} J(\boldsymbol{u})\right\} \tag{14.68}$$

the state and control variables being subjected to the constraints

$$\phi_i = \dot{x}_i - f_i(\boldsymbol{x}, \boldsymbol{u}, t) = 0 \qquad i = 1, \ldots, n \tag{14.69}$$

$$k_j(\boldsymbol{x}(t_0), \boldsymbol{u}(t_0), t_0) = 0 \qquad j = 1, \ldots, n_0 \tag{14.70}$$

$$l_j(\boldsymbol{x}(t_f), \boldsymbol{u}(t_f), t_f) = 0 \qquad j = n_0 + 1, \ldots, n_0 + n_1 \leq 2n + 2 \tag{14.71}$$

In the criterion (14.67), the first term G is called the algebraic part and the second term is called the integral part. Note that the Eqs. (14.69) represent the dynamic model of the process.

When the state can only vary in a given domain, it is preferable to introduce new variables. For example, in the case of a one-dimensional state x, such that $a \leq x \leq b$, we can set the variable z such that $(x-a)(b-x) = z^2$. It is also possible to do the same for the control u when the latter is bounded between two minimum and maximum values.

14.4.2 Variation of the Criterion

In the most general case, the criterion variation is equal to

$$\begin{aligned}\delta J = & \int_{t_0}^{t_f} \left[\left(\frac{\partial F}{\partial \boldsymbol{x}} \right)^T \delta \boldsymbol{x} + \left(\frac{\partial F}{\partial \boldsymbol{u}} \right)^T \delta \boldsymbol{u} \right] dt + F(\boldsymbol{x}_f, \boldsymbol{u}_f, t_f)\, \delta t_f - F(\boldsymbol{x}_0, \boldsymbol{u}_0, t_0)\, \delta t_0 \\ & + \left[\left(\frac{\partial G}{\partial t_0} \right) \delta t_0 + \left(\frac{\partial G}{\partial \boldsymbol{x}_0} \right) \delta \boldsymbol{x}_0 + \left(\frac{\partial G}{\partial \boldsymbol{u}_0} \right) \delta \boldsymbol{u}_0 \right] \\ & + \left[\left(\frac{\partial G}{\partial t_f} \right) \delta t_f + \left(\frac{\partial G}{\partial \boldsymbol{x}_f} \right) \delta \boldsymbol{x}_f + \left(\frac{\partial G}{\partial \boldsymbol{u}_f} \right) \delta \boldsymbol{u}_f \right]\end{aligned} \tag{14.72}$$

The derivative of variations $\delta \boldsymbol{x}$ can be expressed from the state equations as

$$\frac{d}{dt}\delta \boldsymbol{x} = \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \right) \delta \boldsymbol{x} + \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \right) \delta \boldsymbol{u} \quad \Longrightarrow \quad \frac{d}{dt}\delta \boldsymbol{x} - \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \right) \delta \boldsymbol{x} - \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \right) \delta \boldsymbol{u} = 0 \tag{14.73}$$

The latter equation can be multiplied by the Lagrange multipliers and integrated between t_0 and t_f, so that

$$\int_{t_0}^{t_f} \boldsymbol{\psi}(t)^T \left\{ \frac{d}{dt}\delta \boldsymbol{x} - \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \right) \delta \boldsymbol{x} - \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \right) \delta \boldsymbol{u} \right\} dt = 0 \tag{14.74}$$

By summing this equation and Eq. (14.72), one obtains

$$\begin{aligned}\delta J = & \int_{t_0}^{t_f} \left\{ \left[\left(\frac{\partial F}{\partial \boldsymbol{x}} \right)^T - \boldsymbol{\psi}(t)^T \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \right] \delta \boldsymbol{x} + \left[\left(\frac{\partial F}{\partial \boldsymbol{u}} \right)^T - \boldsymbol{\psi}(t)^T \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \right] \delta \boldsymbol{u} \right\} dt \\ & + F(\boldsymbol{x}_f, \boldsymbol{u}_f, t_f)\, \delta t_f - F(\boldsymbol{x}_0, \boldsymbol{u}_0, t_0)\, \delta t_0 \\ & + \left[\left(\frac{\partial G}{\partial t_0} \right) \delta t_0 + \left(\frac{\partial G}{\partial \boldsymbol{x}_0} \right) \delta \boldsymbol{x}_0 + \left(\frac{\partial G}{\partial \boldsymbol{u}_0} \right) \delta \boldsymbol{u}_0 \right] \\ & + \left[\left(\frac{\partial G}{\partial t_f} \right) \delta t_f + \left(\frac{\partial G}{\partial \boldsymbol{x}_f} \right) \delta \boldsymbol{x}_f + \left(\frac{\partial G}{\partial \boldsymbol{u}_f} \right) \delta \boldsymbol{u}_f \right] \\ & + \int_{t_0}^{t_f} \boldsymbol{\psi}(t)^T \frac{d}{dt}\delta \boldsymbol{x}\, dt\end{aligned} \tag{14.75}$$

The last integral of (14.75) can be integrated into parts, thus

$$\int_{t_0}^{t_f} \boldsymbol{\psi}(t)^T \frac{d}{dt}\delta\boldsymbol{x}\, dt = \boldsymbol{\psi}(t_f)^T \delta\boldsymbol{x}_f - \boldsymbol{\psi}(t_0)^T \delta\boldsymbol{x}_0 - \int_{t_0}^{t_f} \dot{\boldsymbol{\psi}}(t)^T\, \delta\boldsymbol{x}\, dt \tag{14.76}$$

If, according to the Hamilton-Jacobi formalism (Sect. 14.4.5), we furthermore introduce the Hamiltonian H equal to

$$H(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\psi}, t) = -F(\boldsymbol{x}, \boldsymbol{u}, t) + \boldsymbol{\psi}(t)^T\, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) \tag{14.77}$$

the variation of the criterion becomes

$$\begin{aligned}
\delta J = & \int_{t_0}^{t_f} \left\{ -\left[\left(\frac{\partial H}{\partial \boldsymbol{x}} \right)^T + \dot{\boldsymbol{\psi}}(t)^T \right] \delta\boldsymbol{x} - \left[\left(\frac{\partial H}{\partial \boldsymbol{u}} \right)^T \right] \delta\boldsymbol{u} \right\} dt \\
& + \left[\left(\frac{\partial G}{\partial t_0} \right) + H(\boldsymbol{x}_0, \boldsymbol{u}_0, \boldsymbol{\psi}_0, t_0) \right] \delta t_0 + \left[\left(\frac{\partial G}{\partial \boldsymbol{x}_0} \right) - \boldsymbol{\psi}(t_0)^T \right] \delta\boldsymbol{x}_0 + \left(\frac{\partial G}{\partial \boldsymbol{u}_0} \right) \delta\boldsymbol{u}_0 \\
& + \left[\left(\frac{\partial G}{\partial t_f} \right) - H(\boldsymbol{x}_f, \boldsymbol{u}_f, \boldsymbol{\psi}_f, t_f) \right] \delta t_f + \left[\left(\frac{\partial G}{\partial \boldsymbol{x}_f} \right) + \boldsymbol{\psi}(t_f)^T \right] \delta\boldsymbol{x}_f + \left(\frac{\partial G}{\partial \boldsymbol{u}_f} \right) \delta\boldsymbol{u}_f
\end{aligned} \tag{14.78}$$

This equation, giving the variation of the criterion, is necessary for understanding the origin of Hamilton-Jacobi equations (Sect. 14.4.5).

14.4.3 Euler Conditions

According to the performance index, the augmented function F^a is defined

$$F^a(\boldsymbol{x}, \dot{\boldsymbol{x}}, \boldsymbol{u}, t) = F(\boldsymbol{x}, \boldsymbol{u}, t) + \sum_{i=1}^{n} \lambda_i\, \phi_i \tag{14.79}$$

Notice that the function G does not intervene in this augmented function, as G depends only on the terminal conditions. G would only intervene in F^a if the terminal conditions were varying.

The variables are the control vector $u(t)$, the state vector $x(t)$ and the Euler-Lagrange multipliers λ. Euler conditions give

$$\begin{array}{lll}
\dfrac{\partial F^a}{\partial u_j} - \dfrac{d}{dt}\dfrac{\partial F^a}{\partial \dot{u}_j} = 0 & & j = 1, \ldots, m \\
\dfrac{\partial F^a}{\partial x_i} - \dfrac{d}{dt}\dfrac{\partial F^a}{\partial \dot{x}_i} = 0 & & i = 1, \ldots, n \\
\dfrac{\partial F^a}{\partial \lambda_i} - \dfrac{d}{dt}\dfrac{\partial F^a}{\partial \dot{\lambda}_i} = 0 & & i = 1, \ldots, n
\end{array} \tag{14.80}$$

The third group of this system of equations corresponds to the constraints $\phi_i = 0$, thus is a system of differential equations with respect to $\dot{x}$. The first group is a system of algebraic equations. The second group is a system of differential equations with respect to $\dot{\lambda}$.

If inequality constraints of the type (14.3) or (14.4) are present, the Valentine's method already discussed in Eqs. (14.12) and (14.13) should be used to modify F^a consequently.

On the other hand, the terminal conditions (14.70) and (14.71), which are transversality and discontinuity conditions, as well as the conditions relative to the second variations will have to be verified.

The transversality equations deduced from Eqs. (14.43) and (14.44) are: At the initial time t_0

$$\begin{aligned} &\left[-\frac{\partial G}{\partial t_0}+\left(F^a-\boldsymbol{\lambda}^T\,\dot{\boldsymbol{x}}\right)_0\right]\delta t_0+\left[-\frac{\partial G}{\partial \boldsymbol{x_0}}+\boldsymbol{\lambda}(t_0)\right]^T\delta \boldsymbol{x}_0=0\\ &\text{with:}\quad \left(\frac{\partial \boldsymbol{k}}{\partial t}\right)_0\delta t_0+\left(\frac{\partial \boldsymbol{k}}{\partial \boldsymbol{x}}\right)_0\delta \boldsymbol{x}_0=0 \end{aligned} \tag{14.81}$$

At the final time t_f

$$\begin{aligned} &\left[\frac{\partial G}{\partial t_f}+\left(F^a-\boldsymbol{\lambda}^T\,\dot{\boldsymbol{x}}\right)_f\right]\delta t_f+\left[\frac{\partial G}{\partial \boldsymbol{x_f}}+\boldsymbol{\lambda}(t_f)\right]^T\delta \boldsymbol{x}_f=0\\ &\text{with:}\quad \left(\frac{\partial \boldsymbol{l}}{\partial t}\right)_f\delta t_f+\left(\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{x}}\right)_f\delta \boldsymbol{x}_f=0 \end{aligned} \tag{14.82}$$

At a fixed final time, from Eq. (14.82), the following condition results

$$\boldsymbol{\lambda}(t_f)=-\frac{\partial G}{\partial \boldsymbol{x_f}} \tag{14.83}$$

Example 14.1: **Linear Quadratic Control using Euler Conditions**

A one-dimensional first-order linear system is defined by the following differential equation

$$\dot{x}=ax+bu \tag{14.84}$$

We wish to minimize with respect to control $u(t)$ the quadratic criterion

$$J=\frac{1}{2}\int_0^{t_f}(u^2+x^2)\,dt \tag{14.85}$$

This corresponds to a single-input single-output continuous linear quadratic control with a finite horizon t_f.

The augmented function is defined

$$F^a=\frac{1}{2}\left(u^2+x^2\right)+\lambda(\dot{x}-ax-bu) \tag{14.86}$$

The Euler conditions for optimality give

$$\begin{aligned} &\frac{\partial F^a}{\partial u}-\frac{d}{dt}\frac{\partial F^a}{\partial \dot{u}} &&=u-\lambda b=0\\ &\frac{\partial F^a}{\partial x}-\frac{d}{dt}\frac{\partial F^a}{\partial \dot{x}} &&=x-\lambda a-\dot{\lambda}=0 \qquad \text{with:}\quad \lambda(t_f)=-\frac{\partial G}{\partial x_f}=0\\ &\frac{\partial F^a}{\partial \lambda}-\frac{d}{dt}\frac{\partial F^a}{\partial \dot{\lambda}} &&=\dot{x}-ax-bu=0 \qquad \text{with:}\quad x(0)=x_0 \end{aligned} \tag{14.87}$$

that is, two differential equations and an algebraic equation. By replacing λ with respect to u from the first equation into the second one, and by eliminating u by differentiating the third equation, we obtain a unique differential equation with respect to the state

$$\ddot{x} - (a^2 + b^2)x = 0 \tag{14.88}$$

hence the general solution

$$x(t) = \alpha \exp(-\sqrt{a^2 + b^2}\, t) + \beta \exp(\sqrt{a^2 + b^2}\, t) \tag{14.89}$$

α and β are two constants to be determined from the terminal condition on the state $x(0)$ and the transversality condition on the adjoint variable $\lambda(t_f)$. However, this is a two-boundary problem that is difficult to solve. The control $u(t)$ is deduced from the model Eq. (14.84).

Numerical Issue:

Indeed, a possibility for obtaining the numerical solution is the following. Given an initial value of the control vector, the differential Eq. (14.84) describing the model is integrated forwards in time, then the adjoint variable differential Eq. (14.87) is integrated backwards in time on the basis of the previous states. A new control profile is obtained, for example, by a gradient method, and the process is repeated iteratively until there is practically no change in the profiles or the criterion is no more improved.

14.4.4 Weierstrass Condition and Hamiltonian Maximization

The Weierstrass condition (14.47) relative to second variations is applied to the augmented function

$$F^a(\boldsymbol{x}, \dot{\boldsymbol{x}}, \boldsymbol{u}, t) = F(\boldsymbol{x}, \boldsymbol{u}, t) + \boldsymbol{\lambda}^T \left[\dot{\boldsymbol{x}} - \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t)\right] \tag{14.90}$$

in the neighbourhood of the optimum, thus

$$F^a(\boldsymbol{x}^*, \dot{\boldsymbol{x}}, \boldsymbol{u}, t) - F^a(\boldsymbol{x}^*, \dot{\boldsymbol{x}}^*, \boldsymbol{u}^*, t) - (\dot{\boldsymbol{x}} - \dot{\boldsymbol{x}}^*)^T \left(\frac{\partial F^a}{\partial \dot{\boldsymbol{x}}}\right)_* \geq 0 \tag{14.91}$$

By clarifying these terms and using the constraints

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}, t) \\ \dot{\boldsymbol{x}}^* &= \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}^*, t) \end{aligned} \tag{14.92}$$

the Weierstrass condition is simplified as

$$\begin{aligned} &F(\boldsymbol{x}^*, \boldsymbol{u}, t) - F(\boldsymbol{x}^*, \boldsymbol{u}^*, t) - \boldsymbol{\lambda}^T \left(\boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}, t) - \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}^*, t)\right) \geq 0 \Longleftrightarrow \\ &\left[\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}^*, t) - F(\boldsymbol{x}^*, \boldsymbol{u}^*, t)\right] - \left[\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}, t) - F(\boldsymbol{x}^*, \boldsymbol{u}, t)\right] \geq 0 \end{aligned} \tag{14.93}$$

in which we recognize the expression of the Hamiltonian (setting $\boldsymbol{\lambda} = \boldsymbol{\psi}$)

$$H(\boldsymbol{x}^*, \boldsymbol{u}, \boldsymbol{\lambda}, t) = -F(\boldsymbol{x}^*, \boldsymbol{u}, t) + \boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}, t) \tag{14.94}$$

We then obtain the fundamental result that the optimal control maximizes the Hamiltonian while respecting the constraints

$$H(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{\lambda}, t) \geq H(\boldsymbol{x}^*, \boldsymbol{u}, \boldsymbol{\lambda}, t) \tag{14.95}$$

which will be generalized as Pontryagin's maximum principle.

Legendre-Clebsch condition for small variations would have allowed us to obtain the stationarity condition at the optimal trajectory, in the absence of constraints, as

$$\left(\frac{\partial H}{\partial u}\right)_* = 0 \tag{14.96}$$

and

$$\left(\frac{\partial^2 H}{\partial u^2}\right)_* \leq 0 \tag{14.97}$$

14.4.5 Hamilton-Jacobi Conditions and Equation

The Hamiltonian[3] is deduced from the criterion (14.67) and from constraints (14.69); it is equal to

$$H(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\psi}(t), t) = -F(\boldsymbol{x}(t), \boldsymbol{u}(t), t) + \boldsymbol{\psi}^T(t)\, \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), t) \tag{14.98}$$

The variation of the criterion has been expressed with respect to the Hamiltonian through Eq. (14.78). We deduce the canonical system of Hamilton conditions

$$\begin{aligned} \dot{\boldsymbol{x}} &= H_{\boldsymbol{\psi}} \\ \dot{\boldsymbol{\psi}} &= -H_{\boldsymbol{x}} \end{aligned} \tag{14.99}$$

which are equivalent to Euler conditions, to which the following equation must be added

$$H_t = -F_t \tag{14.100}$$

The second equation of (14.99) is, in fact, a system of equations called the costate equations, and $\boldsymbol{\psi}$ is called the costate or the vector of adjoint variables.

The derivative of the Hamiltonian is equal to

$$\frac{dH}{dt} = H_{\boldsymbol{x}}^T\, \dot{\boldsymbol{x}} + H_{\boldsymbol{u}}^T\, \dot{\boldsymbol{u}} + H_{\boldsymbol{\psi}}^T\, \dot{\boldsymbol{\psi}} + H_t = H_{\boldsymbol{u}}^T\, \dot{\boldsymbol{u}} + H_t \tag{14.101}$$

[3]Other authors use the definition of the Hamiltonian with an opposite sign before the functional, i.e.

$$H(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\psi}(t), t) = F(\boldsymbol{x}(t), \boldsymbol{u}(t), t) + \boldsymbol{\psi}^T(t)\, \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), t)$$

which changes nothing, as long as we remain at the level of first-order conditions. However, the sign changes in condition (14.83). See also the footnote in Sect. 14.4.6.

If $u(t)$ is an optimal control, one deduces

$$\dot{H} = H_t \tag{14.102}$$

Generally, the concerned physical system is time-invariant so that time does not intervene explicitly in $\boldsymbol{f}$ and also in the functional F, and that Eq. (14.102) becomes

$$\dot{H} = 0 \tag{14.103}$$

In this case, the Hamiltonian is constant along the optimal trajectory.

The transversality conditions are deduced from Eqs. (14.63) and (14.65) or from Eq. (14.78):
At initial time t_0

$$\begin{aligned} &\left[\frac{\partial G}{\partial t_0} + H(t_0)\right] \delta t_0 + \left[\frac{\partial G}{\partial \boldsymbol{x_0}} - \boldsymbol{\psi}(t_0)\right]^T \delta \boldsymbol{x}_0 + \frac{\partial G}{\partial u_0}\, \delta u_0 = 0 \\ &\text{with:} \quad \left(\frac{\partial \boldsymbol{k}}{\partial t}\right)_0 \delta t_0 + \left(\frac{\partial \boldsymbol{k}}{\partial \boldsymbol{x}}\right)_0 \delta \boldsymbol{x}_0 = 0 \end{aligned} \tag{14.104}$$

At final time t_f

$$\begin{aligned} &\left[\frac{\partial G}{\partial t_f} - H(t_f)\right] \delta t_f + \left[\frac{\partial G}{\partial \boldsymbol{x_f}} + \boldsymbol{\psi}(t_f)\right]^T \delta \boldsymbol{x}_f + \frac{\partial G}{\partial u_f}\, \delta u_f = 0 \\ &\text{with:} \quad \left(\frac{\partial \boldsymbol{l}}{\partial t}\right)_f \delta t_f + \left(\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{x}}\right)_f \delta \boldsymbol{x}_f = 0 \end{aligned} \tag{14.105}$$

It is possible to calculate the variation $\delta \mathcal{J}$ associated with the variation δt and with the trajectory change of $\delta \boldsymbol{x}$, the extremity $\boldsymbol{x}_f$ being fixed, for the criterion $\mathcal{J}$ defined in a similar way to Eq. (14.57) by

$$\mathcal{J}(\boldsymbol{x}^*, t) = G(\boldsymbol{x}^*(t_f), t_f) + \int_t^{t_f} F(\boldsymbol{x}^*, \boldsymbol{u}^*, \tau) d\tau \tag{14.106}$$

The variation of the criterion can be expressed with respect to the Hamiltonian

$$\begin{aligned} \delta \mathcal{J}(\boldsymbol{x}^*, t) &= \mathcal{J}(\boldsymbol{x}^* + \delta \boldsymbol{x}(t), t + \delta t) - \mathcal{J}(\boldsymbol{x}^*, t) \\ &= -\left(F - F^T_{\dot{\boldsymbol{x}}} \dot{\boldsymbol{x}}^*\right) \delta t - F^T_{\dot{\boldsymbol{x}}} \delta \boldsymbol{x}(t) \\ &= -F(\boldsymbol{x}^*, \boldsymbol{u}^*, t)\, \delta t \\ &= \left[H(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{\psi}, t) - \boldsymbol{\psi}^T(t)\, \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}^*, t)\right] \delta t \\ &= H(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{\psi}, t)\, \delta t - \boldsymbol{\psi}^T(t)\, \delta \boldsymbol{x}(t) \end{aligned} \tag{14.107}$$

The optimal control corresponds to a maximum of the Hamiltonian. Frequently, the control vector is bounded in a domain U defined by $\boldsymbol{u}_{\min}$ and $\boldsymbol{u}_{\max}$. In this case, the condition that the Hamiltonian is maximum can be expressed in two different ways:

- When a constraint u_i is reached, the function H defined by Eq. (14.98) must be a maximum.

- When the control is located strictly inside the feasible domain U defined by $\boldsymbol{u}_{\min}$ and $\boldsymbol{u}_{\max}$, the derivative of function H defined by Eq. (14.98) with respect to $\boldsymbol{u}$ is zero

$$\frac{\partial H}{\partial u} = 0 \tag{14.108}$$

This equation provides an implicit equation that allows us to express the optimal control with respect only to variables $\boldsymbol{x}, \boldsymbol{\psi}, t$: $\boldsymbol{u}^* = \boldsymbol{u}^*(\boldsymbol{x}, \boldsymbol{\psi}, t)$, hence the new expression of the criterion

$$\begin{aligned} \delta \mathcal{J}(\boldsymbol{x}^*, t) &= H(\boldsymbol{x}^*, \boldsymbol{u}^*(\boldsymbol{x}, \boldsymbol{\psi}, t), \boldsymbol{\psi}, t)\delta t - \boldsymbol{\psi}^T(t)\, \delta \boldsymbol{x}(t) \\ &= \mathcal{J}_t \delta t + \mathcal{J}_{\boldsymbol{x}}^T \delta \boldsymbol{x} \end{aligned} \tag{14.109}$$

thus by identification

$$\begin{aligned} \mathcal{J}_t &= H(\boldsymbol{x}^*, \boldsymbol{u}^*(\boldsymbol{x}, \boldsymbol{\psi}, t), \boldsymbol{\psi}, t) \\ \mathcal{J}_{\boldsymbol{x}} &= -\boldsymbol{\psi}(t) \end{aligned} \tag{14.110}$$

This equation shows that the optimal value of the Hamiltonian is equal to the derivative of criterion (14.106) with respect to time. The Hamilton-Jacobi equation results

$$\mathcal{J}_t - H(\boldsymbol{x}^*, \boldsymbol{u}^*(\boldsymbol{x}, -\mathcal{J}_{\boldsymbol{x}}, t), -\mathcal{J}_{\boldsymbol{x}}, t) = 0 \tag{14.111}$$

with boundary condition

$$\mathcal{J}(\boldsymbol{x}_f^*, t_f) = G(\boldsymbol{x}^*(t_f), t_f) \tag{14.112}$$

The Hamilton-Jacobi equation is a first-order partial derivative equation with respect to the sought function $\mathcal{J}$. Its solving is, in general, analytically impossible for a nonlinear system. In the case of a linear system such as Eq. (14.214), its solving is possible and leads to a Riccati differential equation (14.231). Thus, it is possible to calculate the optimal control law by state feedback. Recall that the Hamilton-Jacobi Eq. (14.111) in discrete form corresponds to the Bellman optimality principle in dynamic programming (Sect. 14.5).

14.4.6 Maximum Principle

Now, examine briefly the Maximum Principle[4] (Pontryaguine et al., 1974) about process optimal control. Pontryagin emphasizes several points:

- An important difference with respect to variational methods is that it is not necessary to consider two close controls in the admissible control domain.

[4] In many articles, authors refer to the Minimum Principle, which simply results from the definition of the Hamiltonian H with an opposite sign of the functional. Comparing to definition (14.98), they define their Hamiltonian as

$$H(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\psi}(t), t) = F(\boldsymbol{x}(t), \boldsymbol{u}(t), t) + \boldsymbol{\psi}^T(t)\, \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), t) \tag{14.113}$$

With that definition, the optimal control u^* minimizes the Hamiltonian.

- The control variables u_i are physical, thus they are constrained, e.g. $|u_1| \leq u_{\max}$, and we consider that they belong to a domain U. The admissible controls are piecewise continuous, that is, they are continuous nearly everywhere, except at some instants where they can undergo first-order discontinuities (jump from one value to another).
- Very frequently, the optimal control is composed by piecewise continuous functions: the control jumps from one summit of the polyhedron defined by U to another. These cases of control occupying only extreme positions cannot be solved by classical methods.

The process is described by a system of differential equations

$$\dot{x}^i(t) = f^i(\boldsymbol{x}(t), \boldsymbol{u}(t)) \qquad i = 1, \ldots, n \tag{14.114}$$

We seek an admissible control $\boldsymbol{u}$ that transfers the system from point $\boldsymbol{x}_0$ in the phase space to point $\boldsymbol{x}_f$ and minimizes the criterion

$$J = G(\boldsymbol{x}_0, t_0, \boldsymbol{x}_f, t_f) + \int_{t_0}^{t_f} F(\boldsymbol{x}(t), \boldsymbol{u}(t))\, dt \tag{14.115}$$

To the n coordinates x^i in the phase space, we add the coordinate x^0 defined[5] by

$$x^0 = G(\boldsymbol{x}_0, t_0, \boldsymbol{x}(t), t) + \int_{t_0}^{t} F(\boldsymbol{x}(\tau), \boldsymbol{u}(\tau))\, d\tau \tag{14.116}$$

so that if $\boldsymbol{x} = \boldsymbol{x}_f$ then $x^0(t_f) = J$. Its derivative is equal to

$$\frac{dx^0}{dt} = G_{\boldsymbol{x}}^T \boldsymbol{f} + G_t + F(\boldsymbol{x}(t), \boldsymbol{u}(t)) \tag{14.117}$$

If time occurs explicitly in the terminal conditions (14.2), or if it is not first fixed, we add the coordinate x^{n+1} to the state (Boudarel et al., 1969), such that

$$\begin{aligned} x^{n+1} &= t \\ \dot{x}^{n+1} &= 1 \end{aligned} \tag{14.118}$$

The complete system of differential equations would then have dimension $n+2$. In the following, in order not to make the notations cumbersome, we will only consider stationary problems of dimension $n+1$ in the form

$$\dot{x}^i = f^i(\boldsymbol{x}(t), \boldsymbol{u}(t)) \qquad i = 0, \ldots, n \tag{14.119}$$

by deducing f^0 from Eq. (14.116) by derivation (extended notation $\boldsymbol{f}$).

In the phase space of dimension $n+1$, we define the initial point $\boldsymbol{x}_0$ and a straight line π parallel to the axis x^0 (i.e. the criterion), passing through the final point $\boldsymbol{x}_f$. The optimal control is, among the admissible controls such that

[5] This notation is that of Pontryaguine et al. (1974). The superscript corresponds to the rank i of the coordinate while the subscripts (0 and 1) or (0 and f), according to the authors, are reserved for the terminal conditions.

the solution $\boldsymbol{x}(t)$, having as the initial condition $\boldsymbol{x}_0$, intersects the line π, the one which minimizes the coordinate x^0 at the intersection point with π.

We introduce the costate variables ψ such that

$$\dot{\boldsymbol{\psi}} = -\boldsymbol{f}_{\boldsymbol{x}}^T\,\boldsymbol{\psi} \Longleftrightarrow \dot{\psi}_i = -\sum_{j=0}^{n} \frac{\partial f^j(\boldsymbol{x}(t),\boldsymbol{u}(t))}{\partial x^i}\,\psi_j \qquad i = 0,\ldots,n \tag{14.120}$$

This system admits a unique solution $\boldsymbol{\psi}$ composed of piecewise continuous functions, corresponding to the control $\boldsymbol{u}$ and presenting the same discontinuity points.

We then consider the Hamiltonian to be equal to the scalar product of functions ψ and f

$$H(\boldsymbol{\psi},\boldsymbol{x},\boldsymbol{u}) = \boldsymbol{\psi}^T\,\boldsymbol{f} = \sum_{i=0}^{n} \psi_i\, f^i \qquad i = 0,\ldots,n \tag{14.121}$$

The systems can be written again in the Hamilton canonical form

$$\begin{aligned} \frac{dx^i}{dt} &= \frac{\partial H}{\partial \psi_i} \qquad i = 0,\ldots,n \\ \frac{d\psi_i}{dt} &= -\frac{\partial H}{\partial x^i} \qquad i = 0,\ldots,n \end{aligned} \tag{14.122}$$

When the solutions $\boldsymbol{x}$ and $\boldsymbol{\psi}$ are fixed, the Hamiltonian depends only on the admissible control $\boldsymbol{u}$, hence the notation

$$\mathcal{M}(\boldsymbol{\psi},\boldsymbol{x}) = \sup_{u\in U} H(\boldsymbol{\psi},\boldsymbol{x},\boldsymbol{u}) \tag{14.123}$$

in order to mean that $\mathcal{M}$ is the maximum of H at fixed $\boldsymbol{x}$ and $\boldsymbol{\psi}$, or further

$$H(\boldsymbol{\psi}^*,\boldsymbol{x}^*,\boldsymbol{u}^*) \geq H(\boldsymbol{\psi}^*,\boldsymbol{x}^*,\boldsymbol{u}^* + \delta\boldsymbol{u}) \qquad \forall\,\delta\boldsymbol{u} \tag{14.124}$$

We consider the admissible controls, defined on $[t_0,t_f]$, to be responding to the previous definition: the trajectory $\boldsymbol{x}(t)$ issued from $\boldsymbol{x}_0$ at t_0 intersects the straight line π at t_f. According to Pontryaguine et al. (1974), the first theorem of the Maximum Principle is expressed as:
So that the control $\boldsymbol{u}(t)$ and the trajectory $\boldsymbol{x}(t)$ is optimal, it is necessary that the continuous and nonzero vector, $\boldsymbol{\psi}(t) = [\psi_0(t),\psi_1(t),\ldots,\psi_n(t)]$ satisfying Hamilton canonical system (14.122), is such that:
1. The Hamiltonian $H[\boldsymbol{\psi}(t),\boldsymbol{x}(t),\boldsymbol{u}(t)]$ reaches its maximum at point $\boldsymbol{u} = \boldsymbol{u}(t)$ $\forall t \in [t_0,t_f]$, thus

$$H[\boldsymbol{\psi}(t),\boldsymbol{x}(t),\boldsymbol{u}(t)] = \mathcal{M}[\boldsymbol{\psi}(t),\boldsymbol{x}(t)] \tag{14.125}$$

2. At the end-time t_f, the relations

$$\psi_0(t_f) \leq 0\,; \quad \mathcal{M}[\boldsymbol{\psi}(t_f),\boldsymbol{x}(t_f)] = 0 \tag{14.126}$$

are satisfied.

With Eq. (14.122) and condition (14.125) being verified, the time functions $\psi_0(t)$ and $\mathcal{M}[\boldsymbol{\psi}(t),\boldsymbol{x}(t)]$ are constant. In this case, the relation (14.126) is verified at any instant t included between t_0 and t_f.

14.4.7 Singular Arcs

In optimal control problems, it often occurs for some time intervals that the Maximum Principle does not give an explicit relation between the control and the state and costate variable: this is a singular optimal control problem which yields singular arcs.

Following Lamnabhi-Lagarrigue (1987), an extremal control has a singular arc $[a, b]$ in $[t_0, t_f]$ if and only if $H_u(\boldsymbol{\psi}^*, \boldsymbol{x}^*, \boldsymbol{u}^*) = 0$ and $H_{uu}(\boldsymbol{\psi}^*, \boldsymbol{x}^*, \boldsymbol{u}^*) = 0$, for all $t \in [a, b]$ and whatever ψ^* satisfies the Maximum Principle.

Example 14.2: **Linear Quadratic Control using Hamilton-Jacobi and Pontryagin Methods**

The previous example is considered again, and, first treated in the framework of Hamilton-Jacobi equations and then in the framework of the Maximum Principle.

Minimize the performance index

$$J = \frac{1}{2} \int_0^{t_f} (u^2 + x_1^2)\, dt \tag{14.127}$$

given

$$\dot{x}_1 = f_1 = a x_1 + b u \tag{14.128}$$

Assume that no constraint exists on the control u.
a) In the context of Hamilton-Jacobi equations, the Hamiltonian would be equal to

$$H = -F + \psi_1 f_1 = -\frac{1}{2}(u^2 + x_1^2) + \psi_1 (a x_1 + b u) \tag{14.129}$$

The Hamilton canonical conditions are

$$\left\{ \begin{array}{ll} \dot{\boldsymbol{x}} & = H_{\boldsymbol{\psi}} \\ \dot{\boldsymbol{\psi}} & = -H_{\boldsymbol{x}} \end{array} \right. \implies \left\{ \begin{array}{lll} \dot{x}_1 & = a x_1 + b u & \text{with:} \quad x_1(0) = x_{1_0} \\ \dot{\psi}_1 & = x_1 - a\, \psi_1 & \text{with:} \quad \psi_1(t_f) = -\dfrac{\partial G}{\partial x_f} = 0 \end{array} \right. \tag{14.130}$$

Maximize H with respect to the control u. As the control u is not constrained, it results that

$$\frac{\partial H}{\partial u} = -u + b\, \psi_1 = 0 \tag{14.131}$$

hence the optimal control

$$u = b\, \psi_1 \tag{14.132}$$

From the previous equations, we draw

$$\ddot{x}_1 - (a^2 + b^2) x_1 = 0 \tag{14.133}$$

The initial state $x_1(0)$ (terminal condition) is given and the final adjoint variable $\psi(t_f)$ (transversality condition) is also known. Thus, this is again a two-boundary problem with the same numerical issue previously stressed in the same example treated by Euler conditions.

b) In the context of Pontryagin Maximum Principle, define the coordinate x_0 (corresponding to x^0 of Eq. (14.116) and here denoted in subscript to avoid confusion with the powers in superscript) such that

$$\dot{x}_0 = f_0 = \frac{1}{2}\,(u^2 + x_1^2) \tag{14.134}$$

The Hamiltonian, defined according to the Maximum Principle is equal to

$$H = \psi_0\, f_0 + \psi_1\, f_1 = \frac{1}{2}\,\psi_0\,(u^2 + x_1^2) + \psi_1\,(a x_1 + b u) \tag{14.135}$$

The Hamilton canonical equations are

$$\begin{cases} \dot{\boldsymbol{x}} &= H_{\boldsymbol{\psi}} \\ \dot{\boldsymbol{\psi}} &= -H_{\boldsymbol{x}} \end{cases} \Longrightarrow \begin{cases} \dot{x}_0 &= \dfrac{\partial H}{\partial \psi_0} = \dfrac{1}{2}\,(u^2 + x_1^2) \\ \dot{x}_1 &= \dfrac{\partial H}{\partial \psi_1} = a x_1 + b u \quad \text{with: } x_1(0) = x_{1_0} \\ \dot{\psi}_0 &= -\dfrac{\partial H}{\partial x_0} = 0 \\ \dot{\psi}_1 &= -\dfrac{\partial H}{\partial x_1} = -\psi_0\, x_1 - \psi_1\, a \quad \text{with: } \psi_1(t_f) = -\dfrac{\partial G}{\partial x_f} = 0 \end{cases} \tag{14.136}$$

The first two equations of this system are the state equations. By setting c_1 as a constant (equal to -1 according to the Hamiltonian given by Eq. (14.129)), from the two following equations, we draw

$$\begin{aligned} \psi_0(t) &= c_1 \\ \dot{\psi}_1 &= -\psi_1\, a - c_1 x_1 \end{aligned} \tag{14.137}$$

We notice that $\psi_0(t)$, which concerns the criterion, is constant as already mentioned in the general theory. Maximize H with respect to the control u. As the control u is not constrained, it gives

$$\frac{\partial H}{\partial u} = \psi_0\, u + \psi_1\, b = 0 \tag{14.138}$$

hence the optimal control

$$u = -\frac{\psi_1\, b}{c_1} \tag{14.139}$$

By using the state Eq. (14.128) and Eq. (14.137), we obtain the same differential equation as (14.88)

$$\ddot{x} - (a^2 + b^2)x = 0 \tag{14.140}$$

Indeed, in the framework of dynamic optimization, Eqs. (14.136) will be integrated with the terminal conditions, including the condition at the fixed initial state $x_1(0) = x_{1_0}$ and the transversality condition at the adjoint variable $\psi_1(t_f) = 0$ at the final time. Of course, the numerical difficulties previously evoked concerning Example 14.1 treated by the Euler equations are encountered here.

Example 14.3: **Mimimum-time Problem with Constraints on the Control**

Consider the following classical example of mechanics: a system reduced to a point is described by its position y, its velocity v and its acceleration u; the latter is thus the input which governs the system position. Although this model is outside the traditional scope of chemical engineering, it has many merits. Essentially, it is very simple so that the different cases studied provide analytical solutions which can be easily understood. The transposition to the chemical engineering field is not necessarily immediate but, for example, the problems of minimum time are very similar. Also, the influence of constraints appears clearly and it illustrates perfectly how the Hamilton-Jacobi method is used.

The physical system is described by the following linear second-order model

$$y^{(2)} = u \tag{14.141}$$

By setting: $x_1 = y$ and $x_2 = \dot{y}$, we get the equivalent state-space model

$$\begin{array}{l} \dot{x}_1 = x_2 \\ \dot{x}_2 = u \end{array} \tag{14.142}$$

We wish to go from the state $(0, 0)$ to the state $(A, 0)$ in a minimum time. This corresponds to a variation in amplitude A on the output: $y = x_1$, while imposing that its derivative remains zero when $t \to 0$ and $t \to t_f$. We impose that the control must stay in the interval $[-1, 1]$.

The performance index is thus

$$J = \int_0^{t_f} dt \tag{14.143}$$

We treat this problem with Hamilton-Jacobi equations, and the Hamiltonian is equal to

$$H = -1 + \psi_1 \, x_2 + \psi_2 \, u \tag{14.144}$$

We obtain

$$\dot{\boldsymbol{\psi}} = -H_{\boldsymbol{x}} \qquad \Longrightarrow \qquad \left\{ \begin{array}{l} \dot{\psi}_1 = 0 \\ \dot{\psi}_2 = -\psi_1 \end{array} \right. \tag{14.145}$$

H can be represented with respect to the bounded control u, and we notice that, with this control being linear with respect to u, the maximum of H is reached when u is equal to -1 or $+1$ according to the sign of ψ_2 (Fig. 14.1). It results from Fig. 14.1 that

$$u = \text{sign}(\psi_2) \tag{14.146}$$

hence

$$H = -1 + \psi_1 \, x_2 + |\psi_2| \tag{14.147}$$

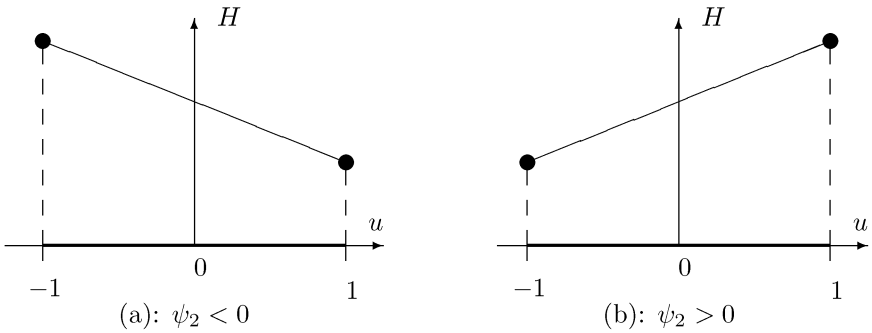


Figure 14.1: Variation of the Hamiltonian with respect to the bounded control ($|u| \leq 1$) and the sign of ψ_2

On the other hand, the Hamilton-Jacobi Eq. (14.111) is

$$\mathcal{J}_t - H(\boldsymbol{x}^*, \boldsymbol{u}^*(\boldsymbol{x}, -\mathcal{J}_{\boldsymbol{x}}, t), -\mathcal{J}_{\boldsymbol{x}}, t) = 0 \tag{14.148}$$

with the criterion $\mathcal{J}$ defined by Eq. (14.106), which verifies

$$\mathcal{J}(\boldsymbol{x}^*, t) = \int_t^{t_f} d\tau \quad \Longrightarrow \quad \mathcal{J}_t = -1 \quad \text{and:} \quad \mathcal{J}_{\boldsymbol{x}} = 0 \tag{14.149}$$

The Hamilton-Jacobi equation results in

$$H(\boldsymbol{x}^*, \boldsymbol{u}^*(\boldsymbol{x}, -\mathcal{J}_{\boldsymbol{x}}, t), -\mathcal{J}_{\boldsymbol{x}}, t) = -1 \quad \Longrightarrow \quad \psi_1 \, x_2^* + |\psi_2| = 0 \tag{14.150}$$

From the differential equations describing the variation of $\boldsymbol{\psi}$, we deduce

$$\begin{aligned} \psi_1 &= c_1 \\ \psi_2 &= -c_1 \, t + c_2 \end{aligned} \tag{14.151}$$

where c_1 and c_2 are constant. As ψ_2 depends linearly on t, ψ_2 crosses the value 0 at the maximum only once at the commutation instant t_c, if the latter exists. The control u thus can take only two values at most, and each value only once

$$\text{if } t \in [0, t_c], u = \pm 1 \quad ; \quad \text{if } t \in]t_c, t_f], u = \mp 1 \tag{14.152}$$

This type of control is called bang-bang. Let $u_0 = \pm 1$. We obtain the solutions of the trajectories

$$\begin{aligned} &\text{if } t \in [0, t_c], u = u_0, \left\{ \begin{array}{l} x_1 = \dfrac{u_0}{2} \, t^2 \\ x_2 = u_0 \, t \end{array} \right. \\ &\text{if } t \in]t_c, t_f], u = -u_0, \left\{ \begin{array}{l} x_1 = -\dfrac{u_0}{2} \, (t - t_f)^2 + A \\ x_2 = -u_0 \, (t - t_f) \end{array} \right. \end{aligned} \tag{14.153}$$

Note that $x_1(t)$ is continuous at the commutation time t_c, so that

$$x_1(t_c^-) = x_1(t_c^+) \quad \left\{ \begin{array}{l} x_1(t_c^-) = \dfrac{u_0}{2} \, t_c^2 \\ x_1(t_c^+) = -\dfrac{u_0}{2} \, (t_c - t_f)^2 + A \end{array} \right. \tag{14.154}$$

which gives

$$(t_c - t_f)^2 + t_c^2 = 2\frac{A}{u_0} \tag{14.155}$$

This equation implies that the right-hand term is positive, so that

$$u_0 = \text{sign}(A) \tag{14.156}$$

Equation (14.155) becomes

$$(t_c - t_f)^2 + t_c^2 = 2|A| \Longrightarrow t_f = t_c + \sqrt{2|A| - t_c^2} \tag{14.157}$$

As the criterion is $J = t_f$, it suffices to minimize the expression of t_f given by Eq. (14.157) with respect to t_c to obtain the commutation time

$$\frac{d(t_c + \sqrt{2|A| - t_c^2})}{dt_c} = 0 \Longrightarrow t_c = \sqrt{|A|} \tag{14.158}$$

In summary, we obtain the three unknowns t_c, t_f, u_0 as

$$\begin{aligned} u_0 &= \text{sign}(A) \\ t_c &= \frac{t_f}{2} \\ t_f &= 2\sqrt{|A|} \approx 2.828 \end{aligned} \tag{14.159}$$

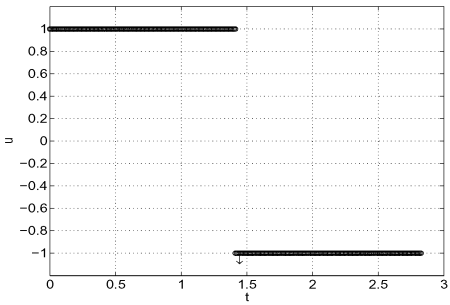


Figure 14.2: Variation of the input

The input presents the bang-bang aspect of Fig. 14.2. The commutation instant is indicated by an arrow. The phase portrait (Fig. 14.3) shows the two arcs and the output trajectory effectively presents a horizontal tangent at the beginning and the end.

This type of problem would correspond to a car driver wanting to go from one point to another in minimum time over a short distance. He first accelerates at the maximum, then brakes at the maximum. In chemical engineering, when the minimum time is searched for an operation, often the solution is that a valve is fully open, then fully closed.

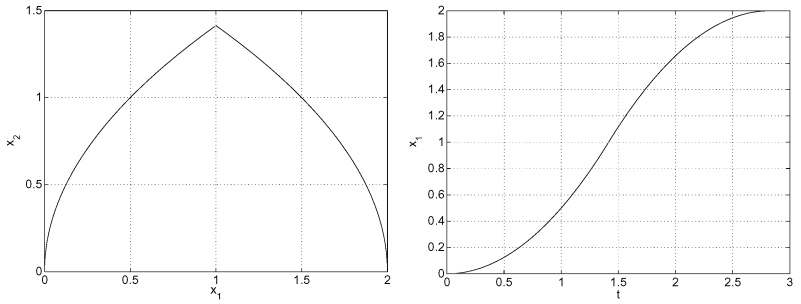


Figure 14.3: Left: phase portrait, right: trajectory of the output (chosen value of the final state x_1: $A = 2$)

Example 14.4: **Miminimum-time Problem with Constraints on the Control and on the State**

Consider the same example as previous but further assume that the state x_2 can be bounded, which amounts to a limit for x_1 in its rate for reaching the final state, thus

$$|x_2| \leq v \qquad \Longrightarrow \qquad x_2^2 - v^2 \leq 0 \tag{14.160}$$

Taking into account the constraint, the Hamiltonian can be written as

$$H = -1 + \psi_1\, x_2 + \psi_2\, u + \mu_1 (x_2^2 - v^2) \tag{14.161}$$

where μ_1 is a Kuhn-Tucker parameter. It results

$$\begin{array}{lll} \mu_1 = 0 & \text{if} & x_2^2 - v^2 < 0 \quad \text{(inside the domain, with respect to } |x_2| < v) \\ \mu_1 \leq 0 & \text{if} & x_2^2 - v^2 = 0 \quad \text{(on the constraint } |x_2| = v) \end{array} \tag{14.162}$$

The condition $\mu_1 \leq 0$ comes from the fact that H must be a maximum. We obtain

$$\left\{ \begin{array}{l} \dot{\psi}_1 = 0 \\ \dot{\psi}_2 = -\psi_1 - 2\,\mu_1\, x_2 \end{array} \right. \tag{14.163}$$

At most, two commutation (or junction) instants t_{c1} and t_{c2} will exist such that

$$t_{c1} = \frac{v}{|u_0|} \qquad ; \quad t_{c2} = t_f - t_{c1} \tag{14.164}$$

Note that the final time t_f is unknown and will be determined by the final condition on the state x_1.

$$\begin{array}{l} \text{if } t \in [0, t_{c1}], u = u_0, \left\{ \begin{array}{l} x_1 = \dfrac{u_0}{2}\, t^2 \\ x_2 = u_0\, t \end{array} \right. \\ \text{if } t \in]t_{c1}, t_{c2}], u = 0, \left\{ \begin{array}{l} x_1 = u_0\, \dfrac{t_{c1}}{2} (2\, t - t_{c1}) \\ x_2 = u_0\, t_{c1} \end{array} \right. \\ \text{if } t \in]t_{c2}, t_f], u = -u_0, \left\{ \begin{array}{l} x_1 = -\dfrac{u_0}{2}\, (t - t_f)^2 + A \\ x_2 = -u_0\, (t - t_f) \end{array} \right. \end{array} \tag{14.165}$$

We deduce

$$t_f = \frac{|u_0|}{u_0\, v}\,(A + \frac{v^2}{u_0}) = \frac{A}{v}\,\mathrm{sign}(u_0) + \frac{v}{|u_0|} \tag{14.166}$$

This final time is equal to 3.12 in the case where the state x_2 is bounded by 0.9 and is larger than the final time without constraint on the state.

The time interval $[t_{c1}, t_{c2}]$ corresponds to a singular arc and is called a singular time interval.

The input (Fig. 14.4) shows first its saturation, then the singular part, then again a saturation with an opposite value. Commutation instants are indicated by arrows. The phase portrait (Fig. 14.5) displays the two commutation times and the saturation for $x_2 = 0.9$. We notice on the trajectory of the output x_1 that during the saturation of x_2 (rate of x_1), the input u being constant, x_1 increases linearly.

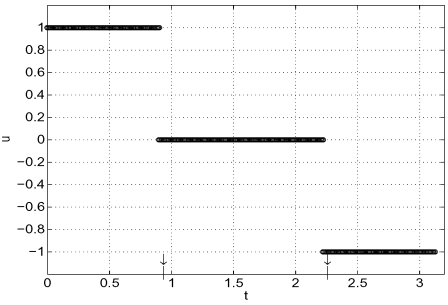


Figure 14.4: Variation of the input

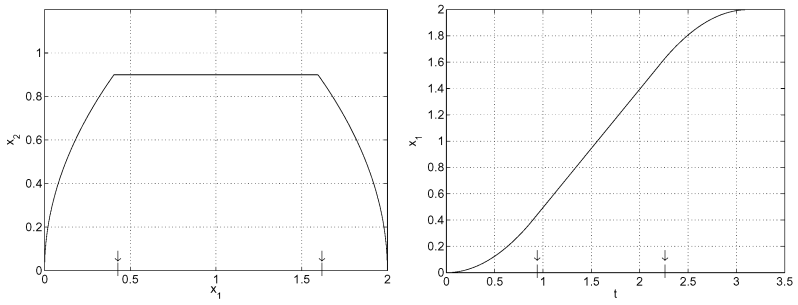


Figure 14.5: Left: phase portrait, right: trajectory of the output (chosen value for the final state x_1: $A = 2$ and the state x_2 is bounded: $|x_2| \leq 0.9$)

14.4.8 Numerical Issues

In general, the dynamic optimization problem results in a set of two systems of first-order ordinary differential equations

$$\begin{array}{lll} \dot{\boldsymbol{x}} & = \dot{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{\psi}, t) & \text{with:} \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \\ \dot{\boldsymbol{\psi}} & = \dot{\boldsymbol{\psi}}(\boldsymbol{x}, \boldsymbol{\psi}, t) & \text{with:} \quad \boldsymbol{\psi}(t_f) = \boldsymbol{\psi}_f \end{array} \tag{14.167}$$

where t_0 and t_f are initial and final time respectively. Thus, it is a two-point boundary-value problem. A criterion J is to be minimized with respect to a control vector. In general, in particular for nonlinear problems, there is no analytical solution.

Different numerical techniques (Bryson, 1999; Latifi et al., 1998) can be used to find the optimal control:

Boundary Condition Iteration:
First, the adjoint vector is initialized and the system (14.167) concerning both $\boldsymbol{x}$ and $\boldsymbol{\psi}$ is integrated. The control vector $u(t)$ results from the maximization of the Hamiltonian with possible constraints on u. The resulting values of $\boldsymbol{\psi}$ at final time t_f are compared with the required values $\boldsymbol{\psi}_f$. From this comparison, new values of the adjoint vector are deduced and the process repeated until convergence.

Multiple Shooting:
This method is very similar to boundary condition iteration except that intermediate points are used in $[t_0, t_f]$ to decompose the problem into a series of boundary condition iteration problems.

Quasi-linearization:
The system (14.167) is linearized around a reference trajectory.

Invariant Embedding:
The initial two-point boundary-value problem is transformed into an initial value problem which is now a system of partial differential equations.

Control Vector Iteration:
Initially, a control vector is assumed. Then, the state equations are integrated forwards in time and then the adjoint equations backwards in time. A gradient method can be used to estimate a new control vector as

$$u_{new}(t) = u_{old}(t) + \alpha \frac{\partial H}{\partial u} \tag{14.168}$$

The choice of the magnitude of displacement α in the gradient direction can be performed by a line search method (Fletcher, 1991) or Rosen gradient projection method(Soeterboek, 1992). In some minimum-time problems such as batch styrene polymerization (Farber and Laurence, 1986), a method based on coordinate transformation can be used, which simply needs that one of the state variables be monotonic such as in general conversion (Kwon and Evans, 1975).

Control Vector Parametrization:
In this method (Goh and Teo, 1988; Teo et al., 1991), the control vector is approximated by basis functions. The control profile can be chosen in different

ways: piecewise constant, piecewise linear or piecewise polynomial, so that the control is expressed as

$$u(t) = \sum a_i \, \phi_i(t) \tag{14.169}$$

where $\phi(t)$ are adequate basis functions, which can be Lagrange polynomials. Thus the optimization is realized with respect to parameters a_i, which govern the control profile.

This method has been used by Fikar et al. (2000) to determine the optimal inputs during the transition from one couple of set points to another for a continuous industrial distillation column. A full model of the distillation column was used by the authors.

Collocation on Finite Elements or Control and State Parameterization: Both state and control variables are approximated by basis functions such as Lagrange polynomials (Biegler, 1984; Cuthrell and Biegler, 1987). The final problem is solved by nonlinear programming.

Consider the following general optimization problem in the form

$$(u, q)^* = \arg\{\min_{u(t),q} J\} \tag{14.170}$$

with the criterion to be minimized

$$J = \Psi[x(t_f), q, t_f] + \int_{t_o}^{tf} \Phi[x(t), u(t), q, t] dt \tag{14.171}$$

subject to the constraints

$$\begin{cases} g[x(t), u(t), q, t] \leq 0 & \text{instantaneous inequality constraints} \\ h[x(t), u(t), q, t] = 0 & \text{instantaneous equality constraints} \\ \dot{x}(t) = f[x(t), u(t), q, t] & \text{state model} \\ x(t_o) = x_o & \text{initial conditions} \\ x_{inf} \leq x(t) \leq x_{sup} & \text{state domain} \\ u_{inf} \leq u(t) \leq u_{sup} & \text{control domain} \\ q_{inf} \leq q \leq q_{sup} & \text{parameter domain} \end{cases} \tag{14.172}$$

where u is the control variables vector, q is the control parameters vector, J is the performance criterion.

Assume that the time domain is divided into subdomains. In this case, an orthogonal collocation is realized at each subdomain $= [\alpha_i, \alpha_{i+1}]$, called a finite element $\Delta\alpha_i$, which is bounded by two knots: α_i and α_{i+1} whose positions are a priori unknown. With these finite elements, the state and control variables are approximated by Lagrange polynomials of orders nc and $nc-1$ respectively

$$x^i_{nc}(\tau) = \sum_{j=0}^{nc} a_{ij} \Phi_{ij}(\tau) \quad ; \quad \Phi_{ij}(\tau) = \prod_{k=0,\neq j}^{nc} \left(\frac{\tau - \tau_{ik}}{\tau_{ij} - \tau_{ik}}\right) \qquad \text{for} \quad i = 1, \ldots, ne \tag{14.173}$$

$$u^i_{nc-1}(\tau) = \sum_{j=1}^{nc} b_{ij} \Psi_{ij}(\tau) \quad ; \quad \Psi_{ij}(\tau) = \prod_{k=1,\neq j}^{nc} \left(\frac{\tau - \tau_{ik}}{\tau_{ij} - \tau_{ik}}\right) \qquad \text{for} \quad i = 1, \ldots, ne \tag{14.174}$$

where ne is the number of finite elements. The dimensionless time, τ, allows to treat easily free end-time problems.

The τ_{ij} are defined as

$$\tau_{ij} = \alpha_i + \gamma_j(\alpha_{i+1} - \alpha_i) \quad ; \quad i = 1, \ldots, ne \quad ; \quad j = 0, \ldots, nc \tag{14.175}$$

where $\gamma_0 = 0$ and the γ_j $(j = 1, \ldots, nc)$ are the zeros of a Legendre polynomial defined on $[0, 1]$.

Replacing the state and control variables by their polynomial approximations in the state system leads to the following algebraic residual equations

$$r(\tau_{il}) = \sum_{j=0}^{nc} a_{ij}\dot{\Phi}_{ij}(\tau_{il}) - t_f.f[a_{il}, b_{il}, q, \tau_{il}] = 0 \quad ; \; l = 1, \ldots, nc \, ; \; i = 1, \ldots, ne \tag{14.176}$$

where

$$\dot{\Phi}_{ij}(\tau_{il}) = \frac{\dot{\Phi}_j(\tau_l)}{\Delta\alpha_i} \tag{14.177}$$

Then, it is necessary to impose the continuity of the state variables between two successive finite elements and to bound the extrapolation of the control variables at both ends of the finite elements, since they are only defined inside each element.

Finally, the problem (14.170) can be approximated as follows

$$(a_{il}, b_{il}, q, \alpha_i, t_f)^* = \arg\{\min_{a_{il}, b_{il}, q, \alpha_i, t_f} J[a_{il}, b_{il}, q, t_f]\} \tag{14.178}$$

subject to

$$\begin{cases} g[a_{il}, b_{il}, q, \tau_{il}] \leq 0 \\ h[a_{il}, b_{il}, q, \tau_{il}] = 0 \\ r(\tau_{il}) = 0 \\ a_{10} = x_o \\ a_{i0} = \sum_{j=0}^{nc} a_{i-1j}\Phi_j(\tau = 1) \\ u_{inf} \leq u^i_{nc-1}(\alpha_i) \leq u_{sup} \\ u_{inf} \leq u^i_{nc-1}(\alpha_{i+1}) \leq u_{sup} \\ x_{inf} \leq a_{il} \leq x_{sup} \\ u_{inf} \leq b_{il} \leq u_{sup} \\ q_{inf} \leq q \leq q_{sup} \end{cases} \tag{14.179}$$

The optimization problem parameters are the state and control variables values at the collocation points defined by Eq. (14.175), the final time, and the position of the knots, which correspond at convergence to the control variable discontinuities. The state and control variables are then completely defined by Eqs. (14.173) and (14.174). This method allows us to easily handle all types of constraints.

The resulting nonlinear problem (14.178) can then be solved using a successive quadratic programming technique, such as that developed by Schittkowski (1985).

The drawback of this technique is that a large number of parameters to be optimized are generated. The advantage is that it does not require the use of Hamiltonian and adjoint equations. However, this method has been successfully used by Gentric et al. (1999) to calculate off-line the optimal temperature profile of a batch emulsion copolymerization reactor.

Iterative Dynamic Programming:

The system is simply described by the dynamic model

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}) \tag{14.180}$$

where the initial state vector $\boldsymbol{x}_0$ is given. A performance index $J(\boldsymbol{x}(t_f))$ is to be minimized, the final time t_f being specified. The optimal control policy is searched by stage to stage optimization, where stages correspond to time subintervals. No gradients are necessary and no auxiliary variables are introduced.

The general principle (Luus, 1996) is the following:

1. Divide the total time interval $[t_0, t_f]$ into P subintervals $[0, t_1], \ldots, [t_{i-1}, t_i]$, etc. of equal length L.
2. Choose an initial estimate of control u for each stage, giving the control policy vector $\boldsymbol{u}$ of dimension P. The control belongs to an initial domain denoted by $\boldsymbol{r}_{in}$. Choose the region contraction factor γ and the region restoration factor η.
3. Choose the number of iterations $j_{\max}$ for each pass and the number of passes $q_{\max}$.
4. Set pass index $q = 1$ and iteration index $j = 1$.
5. Set $\boldsymbol{r}^{(j)} = \eta^q \, \boldsymbol{r}_{in}$.
6. Integrate dynamic model (14.180) to get the state vector $\boldsymbol{x}(k-1)$ at each stage k.
7. From that stage, the principle of dynamic programming is really involved. Beginning at stage P, integrate model (14.180) from $t_f - L$ to t_f for each of the R allowable values of continuous control vector given by

$$\boldsymbol{u}(t) = \boldsymbol{u}(k) + [\boldsymbol{u}(k+1) - \boldsymbol{u}(k)](t - t_k)/L \quad , \qquad \forall k \tag{14.181}$$

with

$$\boldsymbol{u}(P-1) = \boldsymbol{u}^*(P-1)^{(j)} + \boldsymbol{D}_1 \, \boldsymbol{r}^{(j)} \quad , \qquad \boldsymbol{u}(P) = \boldsymbol{u}^*(P)^{(j)} + \boldsymbol{D}_2 \, \boldsymbol{r}^{(j)} \tag{14.182}$$

where the superscript * stands for the best value of the previous iteration and $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices of different random numbers between 0 and 1. The control values should satisfy the constraints $\boldsymbol{u}_{\min} \leq \boldsymbol{u} \leq \boldsymbol{u}_{\max}$. The control values that give the best performance index are retained as $\boldsymbol{u}(P-1)$ and $\boldsymbol{u}(P)$.

8. Proceeding backwards, continue the procedure of step 7 until initial time $t = 0$.
9. Reduce the region for allowable control

$$\boldsymbol{r}^{(j+1)} = \gamma \boldsymbol{r}^{(j)} \tag{14.183}$$

using the best control policy coming from step 8 (denoted by *) as the mean value for the allowable control.
10. Set $j = j + 1$ and go to step 7. Continue until $j < j_{\max}$.
11. Set $q = 1$ and go to step 5. Continue until $q < q_{\max}$.
Variants of this method are given in the following: Banga and Carrasco (1998); Bojkov and Luus (1996); Carrasco and Banga (1997); Luus and Hennessy (1999); Mekarapiruk and Luus (1997). Examples of applications to batch chemical reactors are often cited in these articles.

14.5 Dynamic Programming

14.5.1 Classical Dynamic Programming

Dynamic programming (Bellman, 1957; Bellman and Dreyfus, 1962) has found many applications in chemical engineering (Aris, 1961; Roberts, 1964), in particular for economic optimization problems in refineries, and was frequently developed in the 1960s. Among the typical examples, are the optimization of discontinuous reactors or reactors in series, catalyst replacement or regeneration, the optimization of the counter-current extraction process (Aris et al., 1960), the optimal temperature profile of a tubular chemical reactor (Aris, 1960), the optimization of a cracking reaction (Roberts and Laspe, 1961).

In a different domain, a famous problem is the traveller who, having to go from one point to another, must optimize his travel, which includes the possibility of going through different towns.
Optimality Principle (Bellman, 1957):

A policy is optimal if and only if, whatever the initial state and the initial decision, the decisions remaining to be taken constitute an optimal policy with respect to the state resulting from the first decision.

Because of the principle of continuity, the optimal final value of the criterion is entirely determined by the initial condition and the number of stages. In fact, it is possible to start from any stage, even from the last one. For this reason, Kaufmann and Cruon (1965) express the optimality principle in the following manner:
A policy is optimal if, at a given time, whatever the previous decisions, the decisions remaining to be taken constitute an optimal policy with respect to the result of the previous decisions,
or further,
Any sub-policy (from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$) extracted from an optimal policy (from $\boldsymbol{x}_0$ to $\boldsymbol{x}_N$) is itself optimal from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$.

At first, dynamic programming is discussed in the absence of constraints, which could be terminal constraints, constraints at any time (amplitude constraints) on the state $\boldsymbol{x}$ or on the control u, or inequality constraints. Moreover, we assume the absence of discontinuities.

In fact, as this is a numerical and not analytical solution, these particular cases previously mentioned would pose no problem and could be automatically

considered.

In continuous form, the problem is the following:
Consider the state equation

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}, u) \quad \text{with: } \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{14.184}$$

and the performance index to be minimized

$$J(u) = \int_0^{t_f} r(\boldsymbol{x}, u)dt \tag{14.185}$$

where r represents an income or revenue.

In discrete form, the problem becomes:
Consider the state equation

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + f(\boldsymbol{x}_n, u_n)\Delta t \tag{14.186}$$

with $\Delta t = t_{n+1} - t_n$. The control u_n brings the system from the state $\boldsymbol{x}_n$ to the state $\boldsymbol{x}_{n+1}$ and results in an elementary income $r(\boldsymbol{x}_n, u_n)$ (integrating, in fact, the control period Δt, which will be omitted in the following).

According to the performance index in the integral form, define the performance index or total income at instant N (depending on the initial state $\boldsymbol{x}_0$ and the policy $\mathcal{U}_0^{N-1}$ followed from 0 to $N-1$, bringing from the state $\boldsymbol{x}_0$ to the state $\boldsymbol{x}_N$) as the sum of the elementary incomes $r(\boldsymbol{x}_i, u_i)$

$$J_0 = \sum_{i=0}^{N-1} r(\boldsymbol{x}_i, u_i) \tag{14.187}$$

The values of the initial and final states are known

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0 \; ; \quad \boldsymbol{x}(t_N) = \boldsymbol{x}_N \tag{14.188}$$

If the initial instant is n, note the performance index J_n.

The problem is to find the optimal policy $\mathcal{U}_0^{*,N-1}$ constituted by the succession of controls u_i^* $(i = 0, \ldots, N-1)$ minimizing the performance index J_0. We define the optimal performance index $J^*(\boldsymbol{x}_0, 0)$ as

$$J^*(\boldsymbol{x}_0, 0) = \min_{u_i} J_0 = \min_{u_i} \sum_{i=0}^{N-1} r(\boldsymbol{x}_i, u_i) \tag{14.189}$$

This performance index bears on the totality of the N stages and depends on the starting point $\boldsymbol{x}_0$. In fact, the optimality principle can be applied from any instant n, to which corresponds the optimal performance index $J^*(\boldsymbol{x}_n, n)$

From the optimality principle, the following recurrent algorithm of search of the optimal policy is derived

$$J^*(\boldsymbol{x}_n, n) = \min_{u_n}[r(\boldsymbol{x}_n, u_n) + J^*(\boldsymbol{x}_n + f(\boldsymbol{x}_n, u_n), n+1)] \tag{14.190}$$

which allows us to calculate the series $J^*(\boldsymbol{x}_n, n), J^*(\boldsymbol{x}_{n-1}, n-1), \ldots, J^*(\boldsymbol{x}_0, 0)$ from the final state $\boldsymbol{x}_N$.

If the final state is free, choose $J^*(\boldsymbol{x}_N, N) = 0$. In the case where it is constrained, the last input u^*_{N-1} is calculated so as to satisfy the constraint.

The algorithm (14.190) could be written as

$$\begin{aligned} J^*(\boldsymbol{x}_n, n) &= \min_{u_n}[r(\boldsymbol{x}_n, u_n) + \min_{u_{n+1}}[r(\boldsymbol{x}_{n+1}, u_{n+1}) + J^*(\boldsymbol{x}_{n+2}, n+2)]] \\ &= \min_{u_n}[r(\boldsymbol{x}_n, u_n) + \min_{u_{n+1}}[r(\boldsymbol{x}_{n+1}, u_{n+1}) + \ldots]] \end{aligned} \tag{14.191}$$

However, a difficulty resides frequently in the formulation of a given problem in an adequate form for the solving by means of dynamic programming and, with the actual progress of numerical calculation and nonlinear constrained optimization methods, the latter are nowadays employed more. A variant of dynamic programming (Luss, 1990) called iterative dynamic programming can often provide good results with a lighter computational effort (Luss, 1993; Luss and Bojkov, 1994).

Example 14.5: **Application of Dynamic Programming**

Consider again a particular form of example (14.84) in discrete form. The stable system of dimension 1 is then defined by the following discrete-time state equation

$$x_{k+1} = f(x, u) = 0.5\, x_k + u_k \tag{14.192}$$

and the discrete performance index

$$J = \sum_{k=0}^{3} r(x_k, u_k) = \sum_{k=0}^{3} (x_k^2 + u_k^2) \tag{14.193}$$

Terminal constraints are given

$$\begin{cases} x_0 = 0 \\ x_4 = 1 \end{cases} \tag{14.194}$$

In fact, this is a single-input single-output case of discrete linear quadratic control. This problem has been solved using Maple© by symbolic computation.

We therefore wish to find the control u_k such that the sum of the costs r_k up to time k is minimum when the state goes from x_k to x_{k+1}. At each instant n, we will calculate the control u_n such that the performance index is minimized in recurrent form as

$$J^*(x_n, n) = \min_{u_n}[r(x_n, u_n) + J^*(x_{n+1}, n+1)] \tag{14.195}$$

beginning from the last instant N.
The elementary income is equal to

$$r(x_k, u_k) = x_k^2 + u_k^2 \tag{14.196}$$

The process consists of four stages, thus $N = 4$. Set: $a = 0.5$.

First stage:
The final state is fixed: $x_4 = 1$. It results that

$$x_4 = ax_3 + u_3 = 1 \tag{14.197}$$

which, when joined with the expression (14.192) of the state x_3

$$x_3 = ax_2 + u_2 \tag{14.198}$$

gives the optimal control u_3^*

$$u_3^* = 1 - a(ax_2 + u_2) \tag{14.199}$$

which depends on the value of the state x_2 and the control u_2.
Second stage:
The performance index is transformed by using the expression of x_3 from Eq. (14.192) and the expression (14.199) of the optimal control u_3^*

$$\begin{aligned} J(x_0, 0) &= (x_0^2 + u_0^2) + (x_1^2 + u_1^2) + (x_2^2 + u_2^2) + (x_3^2 + u_3^2) \\ &= (x_0^2 + u_0^2) + (x_1^2 + u_1^2) + (x_2^2 + u_2^2) + (ax_2 + u_2)^2 + [1 - a(ax_2 + u_2)]^2 \\ &= (x_0^2 + u_0^2) + (x_1^2 + u_1^2) + J(x_2, 2) \end{aligned} \tag{14.200}$$

Note that at this instant, the performance index is composed of two parts, one depending on instants from 0 to 1, the other one $J(x_2, 2)$, which is essential depending on the instant 2, thus on the state x_2 and the control u_2. The performance index $J(x_2, 2)$ must be minimum with respect to u_2, which gives the optimal control u_2^*

$$\begin{aligned} &\frac{dJ(x_2, 2)}{du_2} = 0 \Longrightarrow \\ &2u_2 + 2(ax_2 + u_2) - 2a[1 - a(ax_2 + u_2)] = 0 \Longrightarrow \\ &u_2^* = -\frac{a(x_2 - 1 + a^2 x_2)}{2 + a^2} = -0.278x_2 + 0.222 \end{aligned} \tag{14.201}$$

Third stage:
The following is given numerically because of the complexity of analytical expressions. The performance index $J(x_0, 0)$ is transformed by using the expression of x_2 from Eq. (14.192) and the expression (14.201) of the optimal control u_2^*

$$\begin{aligned} J(x_0, 0) &= (x_0^2 + u_0^2) + (x_1^2 + u_1^2) + (0.5x_1 + u_1)^2 + (-0.139x_1 - 0.278u_1 + 0.222)^2 \\ &\quad + (0.111x_1 + 0.222u_1 + 0.222)^2 + (0.889 - 0.556x_1 - 0.111u_1)^2 \\ &= (x_0^2 + u_0^2) + J(x_1, 1) \end{aligned} \tag{14.202}$$

The performance index $J(x_1, 1)$ must be minimum with respect to u_1, which gives the optimal control u_1^*

$$\frac{dJ(x_1, 1)}{du_1} = 0 \Longrightarrow u_1^* = -0.266x_1 + 0.0519 \tag{14.203}$$

Fourth stage:
Finally, the performance index is transformed by using the expression of x_1 from Eq. (14.192) and the expression (14.203) of the optimal control u_1^*

$$\begin{aligned} J(x_0,0) =& (x_0^2+u_0^2)+(0.5x_0+u_0)^2+(-0.133x_0-0.266u_0+0.0519)^2 \\ &+(0.117x_0+0.234u_0+0.0519)^2+(-0.0325x_0-0.0649u_0+0.208)^2 \\ &+(0.0260x_0+0.0519u_0+0.234)^2+(0.883-0.0130x_0-0.0260u_0)^2 \end{aligned} \tag{14.204}$$

This performance index must be minimum with respect to u_0, which gives the optimal control u_0^*

$$\frac{dJ(x_0,0)}{du_0}=0 \Longrightarrow u_0^*=-0.266x_0+0.0122 \tag{14.205}$$

Table 14.1: Succession of the optimal states and controls obtained by dynamic programming

Instant k	State x_k	Control u_k^*
0	0	0.01218
1	0.01218	0.04871
2	0.05479	0.20700
3	0.23440	0.88280
4	1	

By knowing the value of the original state x_0, the succession of inputs and thus of the states is easily calculated in Table 14.1.

14.5.2 Hamilton-Jacobi-Bellman Equation

Given the initial state $\boldsymbol{x}_0$ at time t_0, considering the state $\boldsymbol{x}$ and the control u, the optimal trajectory corresponds to the couple $(\boldsymbol{x}, u)$ such that

$$J^*(\boldsymbol{x}_0,t_0)=\min_{u(t)} J(\boldsymbol{x}_0,u,t_0) \tag{14.206}$$

thus the optimal criterion does not depend on the control u.

In an interval $[t, t+\Delta t]$, the Bellman optimality principle as given in the recurrent Eq. (14.190) can be formulated as

$$J^*(\boldsymbol{x}(t),t)=\min_{u(t)}\left\{\int_t^{t+\Delta t} r(\boldsymbol{x},u,\tau)d\tau+J^*(\boldsymbol{x}(t+\Delta t),t+\Delta t)\right\} \tag{14.207}$$

This can be expressed in continuous form as a Taylor series expansion in the neighbourhood of the state $\boldsymbol{x}(t)$ and time t

$$J^*(\boldsymbol{x}(t),t) = \min_{u(t)} \left\{ \begin{array}{l} r(\boldsymbol{x},u,t)\,\Delta t + J^*(\boldsymbol{x}(t),t) + \dfrac{\partial J^*}{\partial t}\,\Delta t + \\ \left(\dfrac{\partial J^*}{\partial \boldsymbol{x}}\right)^T f(\boldsymbol{x},u,t)\,\Delta t + 0(\Delta t) \end{array} \right\} \tag{14.208}$$

Taking the limit when $\Delta t \to 0$ results in the Hamilton-Jacobi-Bellman equation

$$-\frac{\partial J^*}{\partial t} = \min_{u(t)} \left\{ r(\boldsymbol{x},u,t) + \left(\frac{\partial J^*}{\partial \boldsymbol{x}}\right)^T f(\boldsymbol{x},u,t) \right\} \tag{14.209}$$

As the optimal criterion does not depend on control u, it yields $J^*(\boldsymbol{x}(t_f),t_f) = W(\boldsymbol{x}(t_f))$, which gives the boundary condition for the Hamilton-Jacobi-Bellman Eq. (14.209)

$$J^*(\boldsymbol{x},t_f) = W(\boldsymbol{x}) \quad , \qquad \forall \boldsymbol{x} \tag{14.210}$$

The solution of Eq. (14.209) is the optimal control law

$$u^* = g(\frac{\partial J^*}{\partial \boldsymbol{x}}, \boldsymbol{x}, t) \tag{14.211}$$

which, when introduced into Eq. (14.209) gives

$$-\frac{\partial J^*}{\partial t} = r(\boldsymbol{x},g,t) + \left(\frac{\partial J^*}{\partial \boldsymbol{x}}\right)^T f(\boldsymbol{x},g,t) \tag{14.212}$$

whose solution is $J^*(\boldsymbol{x},t)$ subject to the boundary condition (14.210). Equation (14.212) should be compared to the Hamilton-Jacobi Eq. (14.60). Then, the gradient $\partial J^*/\partial \boldsymbol{x}$ should be calculated and returned in (14.211), which gives the optimal state-feedback control law

$$u^* = g(\frac{\partial J^*}{\partial \boldsymbol{x}}, \boldsymbol{x}, t) = h(\boldsymbol{x},t) \tag{14.213}$$

This corresponds to a closed-loop optimal control law.

14.6 Linear Quadratic Control

Among the numerous publications concerning linear optimal control, are, in particular, the books by Anderson and Moore (1971, 1990); Athans and Falb (1966); Bryson and Ho (1975); Grimble and Johnson (1988a,b); Kirk (1970); Kwakernaak and Sivan (1972); Lewis (1986), and more recently, in robust control Maciejowski (1989). Furthermore, among reference papers, cite Kalman (1960, 1963); Kalman and Bucy (1961). Linear quadratic control is presented here in the previously discussed general framework of optimal control.

14.6.1 Continuous-Time Linear Quadratic Control

In continuous time, the system is represented in the state space by the deterministic linear model

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{A}\,\boldsymbol{x}(t) + \boldsymbol{B}\,\boldsymbol{u}(t) \\ \boldsymbol{y}(t) = \boldsymbol{C}\,\boldsymbol{x}(t) \end{cases} \tag{14.214}$$

where $\boldsymbol{u}$ is the control vector of dimension n_u, $\boldsymbol{x}$ the state vector of dimension n and $\boldsymbol{y}$ the output vector of dimension n_y. $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, are matrices of respective sizes $n \times n, n \times n_u, n_y \times n$.

The control $\boldsymbol{u}$ must minimize the classical quadratic criterion

$$J = 0.5\,\boldsymbol{x}^T(t_f)\,\boldsymbol{Q}_f\,\boldsymbol{x}(t_f) + 0.5\int_{t_0}^{t_f} [\boldsymbol{x}^T(t)\,\boldsymbol{Q}\,\boldsymbol{x}(t) + \boldsymbol{u}^T(t)\,\boldsymbol{R}\,\boldsymbol{u}(t)]dt \tag{14.215}$$

where matrices $\boldsymbol{Q}_f$, $\boldsymbol{Q}$, are symmetrical semipositive definite, whereas $\boldsymbol{R}$ is symmetrical positive definite. This criterion tends to bring the state $\boldsymbol{x}$ towards 0. Other criteria have been derived from the original criterion by replacing the state $\boldsymbol{x}$ by $\boldsymbol{z}$ with $\boldsymbol{z} = \boldsymbol{M}\boldsymbol{x}$, where $\boldsymbol{z}$ represents a linear combination of the states, e.g. a measurement, or the output if $\boldsymbol{M} = \boldsymbol{C}$. The problem is then the regulation of $\boldsymbol{z}$. It is possible to incorporate the tracking of a reference trajectory $\boldsymbol{z}^r = \boldsymbol{M}\boldsymbol{x}^r$ by replacing the state $\boldsymbol{x}$ with the tracking error $(\boldsymbol{z}^r - \boldsymbol{z})$.

Thus, the most general criterion can be considered, which takes into account the different previous cases

$$\begin{aligned} J \quad &= 0.5\,(\boldsymbol{z}^r - \boldsymbol{z})^T(t_f)\,\boldsymbol{Q}_f\,(\boldsymbol{z}^r - \boldsymbol{z})(t_f) \\ &+0.5\int_{t_0}^{t_f} [(\boldsymbol{z}^r - \boldsymbol{z})^T(t)\,\boldsymbol{Q}\,(\boldsymbol{z}^r - \boldsymbol{z})(t) + \boldsymbol{u}^T(t)\,\boldsymbol{R}\,\boldsymbol{u}(t)]dt \end{aligned} \tag{14.216}$$

or

$$\begin{aligned} J \quad &= 0.5\,(\boldsymbol{x}^r - \boldsymbol{x})^T(t_f)\,\boldsymbol{M}^T\,\boldsymbol{Q}_f\,\boldsymbol{M}\,(\boldsymbol{x}^r - \boldsymbol{x})(t_f) \\ &+0.5\int_{t_0}^{t_f} [(\boldsymbol{x}^r - \boldsymbol{x})^T(t)\,\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M}\,(\boldsymbol{x}^r - \boldsymbol{x})(t) + \boldsymbol{u}^T(t)\,\boldsymbol{R}\,\boldsymbol{u}(t)]dt \end{aligned} \tag{14.217}$$

The matrices $\boldsymbol{Q}_f, \boldsymbol{Q}, \boldsymbol{R}$ must have dimensions adapted to the retained criterion.

According to the techniques of variational calculation applied to optimal control, introduce the Hamiltonian as in Eq. (14.98)

$$H = -0.5[(\boldsymbol{x}^r - \boldsymbol{x})^T(t)\,\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M}\,(\boldsymbol{x}^r - \boldsymbol{x})(t) + \boldsymbol{u}^T(t)\,\boldsymbol{R}\,\boldsymbol{u}(t)] + \boldsymbol{\psi}(t)^T\,[\boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{B}\,\boldsymbol{u}] \tag{14.218}$$

The Hamilton canonical Eqs. (14.99) provide, besides the state Eq. (14.214), the derivative of the costate vector

$$-\frac{\partial H}{\partial \boldsymbol{x}} = \dot{\boldsymbol{\psi}}(t) = -\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M}(\boldsymbol{x}^r - \boldsymbol{x}) - \boldsymbol{A}^T\,\boldsymbol{\psi} \tag{14.219}$$

with the final transversality condition (assuming that the state $\boldsymbol{x}_f$ is free, that is, not constrained)

$$\boldsymbol{\psi}(t_f) = \boldsymbol{M}^T\,\boldsymbol{Q}_f\,\boldsymbol{M}\,(\boldsymbol{x}^r - \boldsymbol{x})_f \tag{14.220}$$

Moreover, in the absence of constraints, the condition of maximization of the Hamiltonian with respect to $\boldsymbol{u}$ gives

$$\frac{\partial H}{\partial \boldsymbol{u}} = 0 = -\boldsymbol{R}\,\boldsymbol{u} + \boldsymbol{B}^T\,\boldsymbol{\psi} \tag{14.221}$$

hence the optimal control

$$\boldsymbol{u}(t) = \boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{\psi}(t) \tag{14.222}$$

Note that if a different definition of the Hamiltonian is used with a positive sign before the functional, the previous formula results in a negative sign and the non-diagonal terms of the first square matrix of Eq. (14.223) will have opposite signs as in Eq. (14.235) defining the Hamiltonian matrix. It would suffice to change ψ in $-\psi$.

Gathering all these results, the system to be solved becomes a two-point boundary-value problem

$$\begin{aligned}
&\begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ \dot{\boldsymbol{\psi}}(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B}\,\boldsymbol{R}^{-1}\,\boldsymbol{B}^T \\ \boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M} & -\boldsymbol{A}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{\psi}(t) \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^r(t) \\ 0 \end{bmatrix} \\
&\boldsymbol{x}(t_0) = \boldsymbol{x}_0 \\
&\boldsymbol{\psi}(t_f) = \boldsymbol{M}^T\,\boldsymbol{Q}_f\,\boldsymbol{M}\,(\boldsymbol{x}^r - \boldsymbol{x})_f
\end{aligned} \tag{14.223}$$

Regulation Case: $\boldsymbol{x}^r = 0$

A classical approach consists of introducing the transition matrix corresponding to the previous differential system, which can be partitioned such that

$$\begin{bmatrix} \boldsymbol{x}(\tau) \\ \boldsymbol{\psi}(\tau) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_{xx}(\tau,t) & \boldsymbol{\Phi}_{x\psi}(\tau,t) \\ \boldsymbol{\Phi}_{\psi x}(\tau,t) & \boldsymbol{\Phi}_{\psi\psi}(\tau,t) \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{\psi}(t) \end{bmatrix}, \qquad \tau \in [t, t_f] \tag{14.224}$$

This equation can be used at the final instant $\tau = t_f$ (where $\boldsymbol{\psi}(t_f)$ is known) thus

$$\begin{aligned}
\boldsymbol{x}(t_f) &= \boldsymbol{\Phi}_{xx}(t_f,t)\,\boldsymbol{x}(t) + \boldsymbol{\Phi}_{x\psi}(t_f,t)\,\boldsymbol{\psi}(t) \\
\boldsymbol{\psi}(t_f) &= \boldsymbol{\Phi}_{\psi x}(t_f,t)\,\boldsymbol{x}(t) + \boldsymbol{\Phi}_{\psi\psi}(t_f,t)\,\boldsymbol{\psi}(t)
\end{aligned} \tag{14.225}$$

hence

$$\begin{aligned}
\boldsymbol{\psi}(t) = \quad & -[\boldsymbol{\Phi}_{\psi\psi}(t_f,t) + \boldsymbol{M}^T\,\boldsymbol{Q}_f\,\boldsymbol{M}\boldsymbol{\Phi}_{x\psi}(t_f,t)]^{-1} \\
& [\boldsymbol{\Phi}_{\psi x}(t_f,t) + \boldsymbol{M}^T\,\boldsymbol{Q}_f\,\boldsymbol{M}\boldsymbol{\Phi}_{xx}(t_f,t)]\,\boldsymbol{x}(t)
\end{aligned} \tag{14.226}$$

a relation which can be denoted by

$$\boldsymbol{\psi}(t) = -\boldsymbol{M}^T\,\boldsymbol{S}(t)\,\boldsymbol{M}\,\boldsymbol{x}(t) = -\boldsymbol{P}_c(t)\,\boldsymbol{x}(t) \tag{14.227}$$

both to express the proportionality and to verify the terminal condition (at t_f), $\boldsymbol{M}$ being any constant matrix. The subscript c of $\boldsymbol{P}_c$ means that we are treating

the control problem (to be compared with $\boldsymbol{P}_f$ later used for Kalman filtering). The relation (14.227) is also called a backward sweep solution (Bryson, 1999).

This relation joined to the optimal control expression gives

$$\boldsymbol{u}^*(t) = -\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{M}^T\,\boldsymbol{S}(t)\,\boldsymbol{M}\,\boldsymbol{x}(t) = -\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{P}_c(t)\,\boldsymbol{x}(t) \tag{14.228}$$

which shows that this is a state feedback control. By setting $\boldsymbol{M} = \boldsymbol{I}$, we find the classical formula which equals $\boldsymbol{P}_c(t)$ and $\boldsymbol{S}(t)$

$$\boldsymbol{u}^*(t) = -\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{S}(t)\,\boldsymbol{x}(t) \tag{14.229}$$

In fact, the matrix $\boldsymbol{P}_c(t)$ can be calculated directly. Use the relation

$$\boldsymbol{\psi}(t) = -\boldsymbol{P}_c(t)\,\boldsymbol{x}(t) \tag{14.230}$$

inside the system (14.223). The continuous differential Riccati equation results

$$\begin{aligned} \dot{\boldsymbol{P}}_c(t) \quad &= -\boldsymbol{P}_c(t)\,\boldsymbol{A} - \boldsymbol{A}^T\,\boldsymbol{P}_c(t) + \boldsymbol{P}_c(t)\,\boldsymbol{B}\,\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{P}_c(t) - \boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M} \\ &\text{with:} \quad \boldsymbol{P}_c(t_f) = \boldsymbol{M}^T\,\boldsymbol{Q}_f\,\boldsymbol{M} \end{aligned} \tag{14.231}$$

where the matrix $\boldsymbol{P}_c(t)$ is symmetrical semipositive definite. Knowing the solution of this differential equation, the optimal control law can be calculated

$$\boldsymbol{u}^*(t) = -\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{P}_c(t)\,\boldsymbol{x}(t) = -\boldsymbol{K}_c(t)\,\boldsymbol{x}(t) \tag{14.232}$$

Notice that the differential Riccati equation (14.231), being known by its final condition, can be integrated backwards to deduce $\boldsymbol{P}_c(t_0)$, which will allow us to exploit the optimal control law in relation to the differential system (14.214).

If the horizon t_f is infinite, the control law is

$$\boldsymbol{u}^*(t) = -\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{P}_c\,\boldsymbol{x}(t) \tag{14.233}$$

where the matrix $\boldsymbol{P}_c$ is the solution of the algebraic Riccati equation

$$\boldsymbol{P}_c\,\boldsymbol{A} + \boldsymbol{A}^T\,\boldsymbol{P}_c - \boldsymbol{P}_c\,\boldsymbol{B}\,\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{P}_c + \boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M} = 0 \tag{14.234}$$

which is the steady-state form of the differential Riccati Eq. (14.231).

Different methods have been published to solve Eq. (14.234), which poses serious numerical problems. A solution can be to integrate backwards the differential Riccati equation (14.231) until a stationary solution is obtained. Another solution, which is numerically robust and is based on a Schur decomposition method, is proposed by Arnold and Laub (1984); Laub (1979). Consider the Hamiltonian[6] matrix $\boldsymbol{\mathcal{H}}$

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B}\,\boldsymbol{R}^{-1}\,\boldsymbol{B}^T \\ -\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M} & -\boldsymbol{A}^T \end{bmatrix} \tag{14.235}$$

[6] A matrix $\boldsymbol{A}$ of dimension $(2n \times 2n)$ is called Hamiltonian if $\boldsymbol{J}^{-1}\,\boldsymbol{A}^T\,\boldsymbol{J} = -\boldsymbol{A}$ or $\boldsymbol{J} = -\boldsymbol{A}^{-T}\,\boldsymbol{J}\,\boldsymbol{A}$, where $\boldsymbol{J}$ is equal to: $\begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}$.

An important property (Laub, 1979) of Hamiltonian matrices is that if λ is an eigenvalue of a Hamiltonian matrix, $-\lambda$ is also an eigenvalue with the same multiplicity.

whose eigenvalues and eigenvectors are sought. Order the matrix $\boldsymbol{U}$ of the eigenvectors of the Hamiltonian matrix of dimension $2n \times 2n$ in such a way that the first n columns are the eigenvectors corresponding to the stable eigenvalues (negative real or complex with a negative real part), in the form

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} \\ \boldsymbol{U}_{21} & \boldsymbol{U}_{22} \end{bmatrix} \tag{14.236}$$

where the blocks $\boldsymbol{U}_{ij}$ have dimension $n \times n$. The solution to the algebraic Riccati equation (14.234) is then

$$\boldsymbol{P}_c = \boldsymbol{U}_{21}\, \boldsymbol{U}_{11}^{-1} \tag{14.237}$$

Note that the stationary solution to this problem can also be obtained by iterative methods (Arnold and Laub, 1984).

Tracking Case: $\boldsymbol{x}^r \neq 0$

In the presence of the tracking term $\boldsymbol{x}^r$, the differential system (14.223) is not homogeneous anymore, and it is necessary to add a term to (14.224), as

$$\begin{bmatrix} \boldsymbol{x}(\tau) \\ \boldsymbol{\psi}(\tau) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_{xx}(\tau,t) & \boldsymbol{\Phi}_{x\psi}(\tau,t) \\ \boldsymbol{\Phi}_{\psi x}(\tau,t) & \boldsymbol{\Phi}_{\psi\psi}(\tau,t) \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{\psi}(t) \end{bmatrix} + \begin{bmatrix} \boldsymbol{g}_x(\tau,t) \\ \boldsymbol{g}_\psi(\tau,t) \end{bmatrix} \tag{14.238}$$

In the same manner as previously, at the final instant $\tau = t_f$, we obtain

$$\begin{aligned} \boldsymbol{x}(t_f) &= \boldsymbol{\Phi}_{xx}(t_f,t)\, \boldsymbol{x}(t) + \boldsymbol{\Phi}_{x\psi}(t_f,t)\boldsymbol{\psi}(t) + \boldsymbol{g}_x(t_f,t) \\ \boldsymbol{\psi}(t_f) &= \boldsymbol{\Phi}_{\psi x}(t_f,t)\, \boldsymbol{x}(t) + \boldsymbol{\Phi}_{\psi\psi}(t_f,t)\boldsymbol{\psi}(t) + \boldsymbol{g}_\psi(t_f,t) \end{aligned} \tag{14.239}$$

hence

$$\begin{aligned} \boldsymbol{\psi}(t) = \quad & -[\boldsymbol{\Phi}_{\psi\psi}(t_f,t) + \boldsymbol{M}^T\, \boldsymbol{Q}_f\, \boldsymbol{M}\, \boldsymbol{\Phi}_{x\psi}(t_f,t)]^{-1} \\ & \Big\{ [\boldsymbol{\Phi}_{\psi x}(t_f,t) + \boldsymbol{M}^T\, \boldsymbol{Q}_f\, \boldsymbol{M} \boldsymbol{\Phi}_{xx}(t_f,t)]\, \boldsymbol{x}(t) + \\ & \boldsymbol{M}^T\, \boldsymbol{Q}_f\, \boldsymbol{M}(\boldsymbol{g}_x(t_f,t) - x_{r,f}) + \boldsymbol{g}_\psi(t_f,t) \Big\} \end{aligned} \tag{14.240}$$

giving an expression in the form

$$\boldsymbol{\psi}(t) = -\boldsymbol{P}_c(t)\, \boldsymbol{x}(t) + \boldsymbol{s}(t) \tag{14.241}$$

By introducing this expression in (14.219), we again get the differential Riccati Eq. (14.231) whose matrix $\boldsymbol{P}_c$ is a solution, with the same terminal condition. Moreover, we obtain the differential equation giving the vector $\boldsymbol{s}$

$$\begin{aligned} \dot{\boldsymbol{s}}(t) \quad & = [\boldsymbol{P}_c(t)\, \boldsymbol{B}\, \boldsymbol{R}^{-1}\, \boldsymbol{B}^T - \boldsymbol{A}^T]\, \boldsymbol{s}(t) - \boldsymbol{M}^T\, \boldsymbol{Q}\, \boldsymbol{M}\, \boldsymbol{x}^r \\ & \text{with:} \quad \boldsymbol{s}(t_f) = \boldsymbol{M}^T\, \boldsymbol{Q}_f\, \boldsymbol{M}\, \boldsymbol{x}_f^r \end{aligned} \tag{14.242}$$

This equation is often termed feedforward. Knowing the solutions of this differential equation and of the differential Riccati Eq. (14.231), the optimal control law can be calculated and applied

$$\boldsymbol{u}^*(t) = -\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{P}_c(t)\,\boldsymbol{x}(t) + \boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{s}(t) = \boldsymbol{u}_{fb}(t) + \boldsymbol{u}_{ff}(t) \qquad (14.243)$$

In this form, $\boldsymbol{u}_{fb}(t)$ represents a state feedback control (term in $x(t)$), as the gain $\boldsymbol{K}_c$ depends at each instant on the solution of the Riccati equation, and $\boldsymbol{u}_{ff}(t)$ represents a feedforward control (term in $s(t)$). This structure is visible in Fig. 14.6.

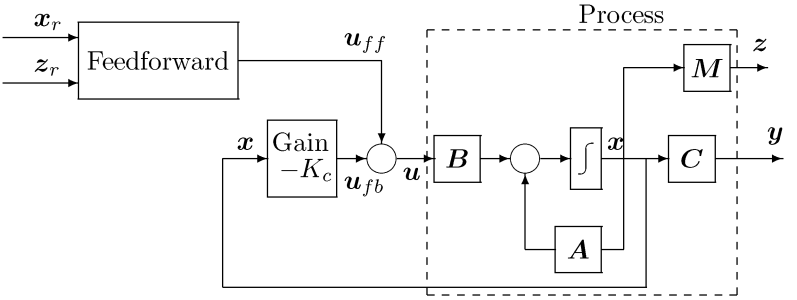


Figure 14.6: Structure of linear quadratic control

In the same manner as in regulation, when the horizon is infinite, $\boldsymbol{P}_c$ is solution of the algebraic Riccati equation (14.234) and $\boldsymbol{s}$ is solution of the algebraic equation

$$\boldsymbol{s} = [\boldsymbol{P}_c\,\boldsymbol{B}\,\boldsymbol{R}^{-1}\,\boldsymbol{B}^T - \boldsymbol{A}^T]^{-1}\,\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M}\,\boldsymbol{x}^r \qquad (14.244)$$

This solution is frequently adopted in actual practice.

Lin (1994) recommends choosing, as a first approach, diagonal criterion weighting matrices with their diagonal terms equal to

$$q_i = 1/(z_i)^2_{\max}\,, \qquad r_i = 1/(u_i)^2_{\max} \qquad (14.245)$$

in order to realize a compromise between the input variations and the performance with respect to the output, while Anderson and Moore (1990) propose taking

$$q_i = 1/\int_0^\infty z_i^2 dt\,, \qquad r_i = 1/\int_0^\infty u_i^2 dt \qquad (14.246)$$

It frequently happens, as in the simple linear example previously treated, that some components of the control vector are bounded

$$|u_i| \leq u_{i,\max} \qquad (14.247)$$

In this case, the optimal control is equal to

$$\boldsymbol{u}^* = \mathrm{sat}(\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{\psi}) \qquad (14.248)$$

defining the saturation function by

$$\text{sat}(u_i) = \begin{cases} u_i & \text{if: } |u_i| \leq u_{i,\max} \\ u_{i,\max} & \text{if: } |u_i| \geq u_{i,\max} \end{cases} \tag{14.249}$$

Example 14.6: **Linear Quadratic Control of an Extractive Distillation Column**

Gilles and Retzbach (1983); Gilles et al. (1980) studied an industrial extractive

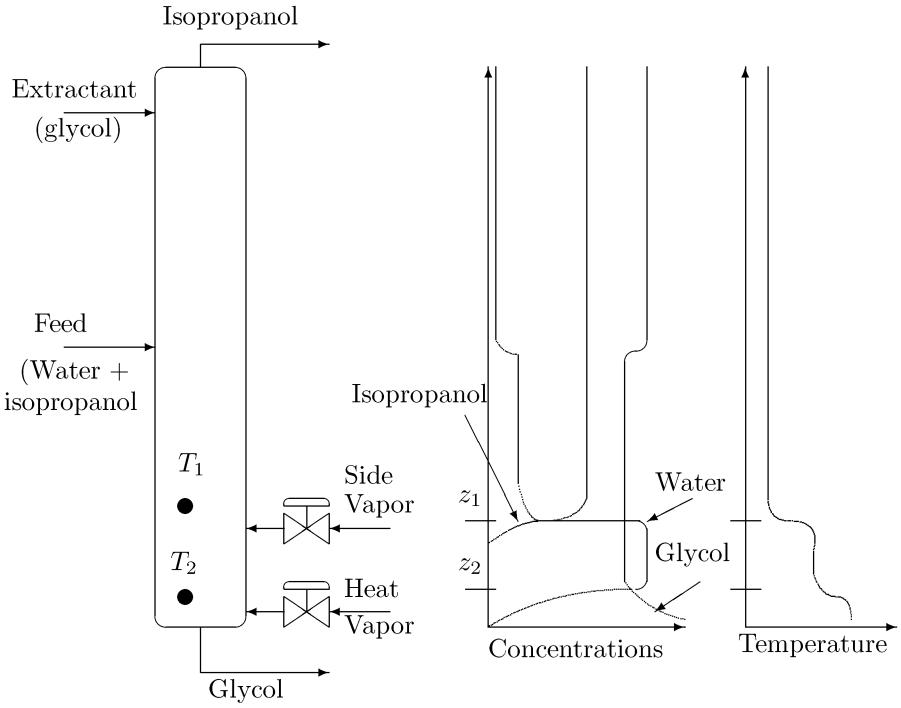


Figure 14.7: Extractive distillation column of Gilles, with concentration and temperature profiles

distillation column (Fig. 14.7) designed to separate water and isopropanol from the feed using glycol as an extracting agent, introduced near the top of the column. The product is purified isopropanol at the top of the column, while glycol leaves at the bottom. The reduced-order model developed by Gilles describes the concentration and temperature profiles in the column. At two places denoted by z_1 and z_2, stiff variations of concentration and temperature occur; at z_1, the interfacial separation between water and isopropanol happens; at z_2, the interfacial separation between water and glycol happens, hence the allure of the profiles (Fig. 14.7). The position of these concentration and temperature fronts depends on the feed flow rate, the heating vapour flow rate and the side vapour flow rate. The feed composition and flow rate are

disturbances, while the heating vapour flow rate and the side vapour flow rate are the manipulated variables.

Denoting by δ the variations with respect to the steady state, the boiler dynamics is represented by the differential equations

$$\begin{aligned} \delta\dot{Q}_1 &= a_{11}\,\delta Q_1 + b_{11}\,\delta u_1 \\ \delta\dot{V}_1 &= a_{21}\,\delta Q_1 + a_{22}\,\delta V_1 \end{aligned} \tag{14.250}$$

with the following notation:
δQ_1: heat flow rate at the boiler,
δV_1: vapour flow rate,
δu_1: heating vapour flow rate.

The variations of the fronts are represented by

$$\begin{aligned} \delta\dot{z}_1 &= b_{32}\,\delta S + f_{31}\,\delta x_{FA1} + f_{32}\,\delta F_A \\ \delta\dot{z}_2 &= b_{42}\,\delta S + f_{42}\,\delta F_A \end{aligned} \tag{14.251}$$

with:
δS: side vapour flow rate,
δx_{FA1}: feed composition,
δF_A: feed flow rate.

The position of the fronts is determined by measuring the temperature by means of thermocouples located in the neighbourhood of the desired positions of the fronts according to a linear law

$$\begin{aligned} \delta T_1 &= c_{13}\,\delta z_1 \\ \delta T_2 &= c_{24}\,\delta z_2 \end{aligned} \tag{14.252}$$

The continuous state-space model of the process is then

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{B}\,\boldsymbol{u} + \boldsymbol{F}\,\boldsymbol{d} \\ \boldsymbol{y} &= \boldsymbol{C}\,\boldsymbol{x} \end{aligned} \tag{14.253}$$

where $\boldsymbol{d}$ represents the disturbances and $\boldsymbol{y}$ the measurements. The vectors are defined with respect to the physical variables by

$$\boldsymbol{x} = \begin{bmatrix} \delta Q_1 \\ \delta V_1 \\ \delta z_1 \\ \delta z_2 \end{bmatrix} \quad ; \quad \boldsymbol{u} = \begin{bmatrix} \delta u_1 \\ \delta S \end{bmatrix} \quad ; \quad \boldsymbol{d} = \begin{bmatrix} \delta x_{FA1} \\ \delta F_A \end{bmatrix} \quad ; \quad \boldsymbol{y} = \begin{bmatrix} \delta T_1 \\ \delta T_2 \end{bmatrix} \tag{14.254}$$

$\boldsymbol{A}$ is a 4×4 matrix, $\boldsymbol{B}$ 4×2, $\boldsymbol{F}$ 4×2, $\boldsymbol{C}$ 2×4. The time unit is hours and temperature unit is degrees Kelvin. Numerical data are the following (all other

elements of the matrices are zero)

$$\begin{array}{llll} a_{11} = -30.3 & b_{11} = 6.15 \times 10^5 & f_{31} = 62.2 & c_{13} = -7.3 \\ a_{21} = 0.12 \times 10^{-3} & b_{32} = 3.04 & f_{32} = 5.76 & c_{24} = -25.0 \\ a_{22} = -6.02 & b_{42} = 0.052 & f_{42} = 5.12 & \\ a_{32} = -3.77 & & & \\ a_{42} = -2.80 & & & \end{array} \tag{14.255}$$

In order to emphasize the influence of the weighting matrices of the quadratic index (14.216), linear quadratic control is employed for different values of the weighting matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$. $\boldsymbol{Q}_f$ is chosen as zero and $\boldsymbol{M}$ is equal to $\boldsymbol{C}$. Two set point step variations have been realized for each output. The set point is perfectly followed (Figs. 14.8 to 14.10). However, note that there is neither noise, nor disturbance. The responsiveness of the tracking increases when the weighting with respect to the control decreases, leading to large variations of the inputs, which could possibly saturate. The coupling of the multivariable system appears clearly at $t = 2\,\mathrm{h}$ and $t = 4\,\mathrm{h}$, when the set point variations concern only one output, respectively y_1 and y_2.

If all the states are considered to be known and $\boldsymbol{M}$ is the identity matrix, the optimized LQ criterion is given by Eq. (14.215). In the case where $\boldsymbol{Q}$ and $\boldsymbol{R}$ are chosen as identity matrices, the solution of the algebraic Riccati equation is then

$$\boldsymbol{P}_c = \begin{bmatrix} 1.626 \times 10^{-6} & 7.388 \times 10^{-7} & 2.778 \times 10^{-8} & -1.626 \times 10^{-6} \\ 7.388 \times 10^{-7} & 3.786 \times 10^{+3} & 1.424 \times 10^{+2} & -8.332 \times 10^{+3} \\ 2.778 \times 10^{-8} & 1.424 \times 10^{+2} & 5.694 \times 10^{+0} & -3.137 \times 10^{+2} \\ -1.626 \times 10^{-6} & -8.332 \times 10^{+3} & -3.137 \times 10^{+2} & 1.834 \times 10^{+4} \end{bmatrix} \tag{14.256}$$

and the corresponding steady-state gain

$$\boldsymbol{K}_c = \begin{bmatrix} 9.999 \times 10^{-1} & 4.544 \times 10^{-1} & 1.709 \times 10^{-2} & -9.999 \times 10^{-1} \\ -8.221 \times 10^{-11} & -4.213 \times 10^{-1} & 9.999 \times 10^{-1} & 1.709 \times 10^{-2} \end{bmatrix} \tag{14.257}$$

These results concerning the algebraic Riccati equation have been obtained by the Hamiltonian method. Note that some well-known control codes may give erroneous solutions for this solving. This system is close to being uncontrollable.

If only the measured states x_3 and x_4 are considered, $\boldsymbol{M}$ is equal to $\boldsymbol{C}$ and the optimized LQ criterion is given by Eq. (14.216). Again, in the case where $\boldsymbol{Q}$ and $\boldsymbol{R}$ are chosen as identity matrices, the solution of the algebraic Riccati equation is then

$$\boldsymbol{P}_c = \begin{bmatrix} 3.292 \times 10^{-12} & 8.531 \times 10^{-7} & -3.269 \times 10^{-7} & -1.593 \times 10^{-6} \\ 8.531 \times 10^{-7} & 2.224 \times 10^{-1} & -9.472 \times 10^{-2} & -4.169 \times 10^{-1} \\ -3.269 \times 10^{-7} & -9.472 \times 10^{-2} & 3.237 \times 10^{-1} & -8.621 \times 10^{-2} \\ -1.593 \times 10^{-6} & -4.169 \times 10^{-1} & -8.621 \times 10^{-2} & 1.174 \times 10^{+0} \end{bmatrix} \tag{14.258}$$

and the corresponding steady-state gain

$$\boldsymbol{K}_c = \begin{bmatrix} 2.025 \times 10^{-6} & 5.247 \times 10^{-1} & -2.011 \times 10^{-1} & -9.796 \times 10^{-1} \\ -1.077 \times 10^{-6} & -3.096 \times 10^{-1} & 9.796 \times 10^{-1} & -2.010 \times 10^{-1} \end{bmatrix} \tag{14.259}$$

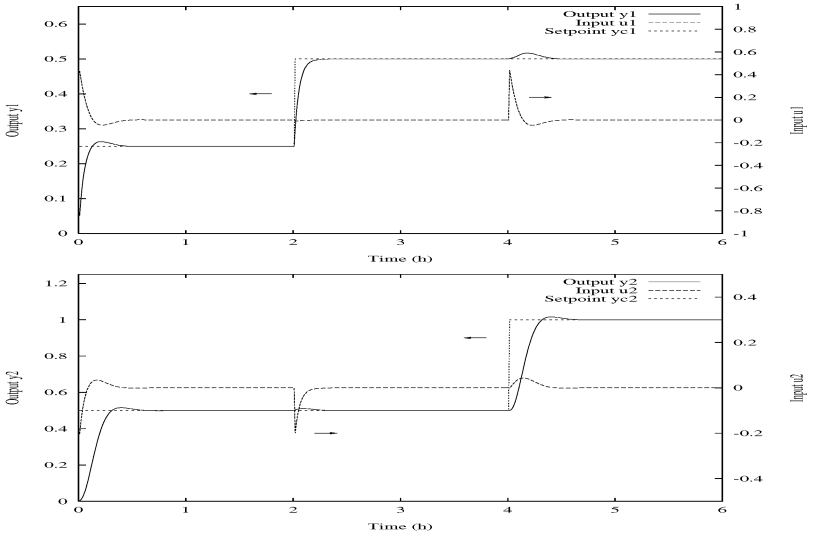


Figure 14.8: Linear quadratic control of Gilles extractive distillation column (criterion weighting: $\boldsymbol{Q} = \boldsymbol{R} = \boldsymbol{I}$). Top: input u_1 and output y_1. Bottom: input u_2 and output y_2

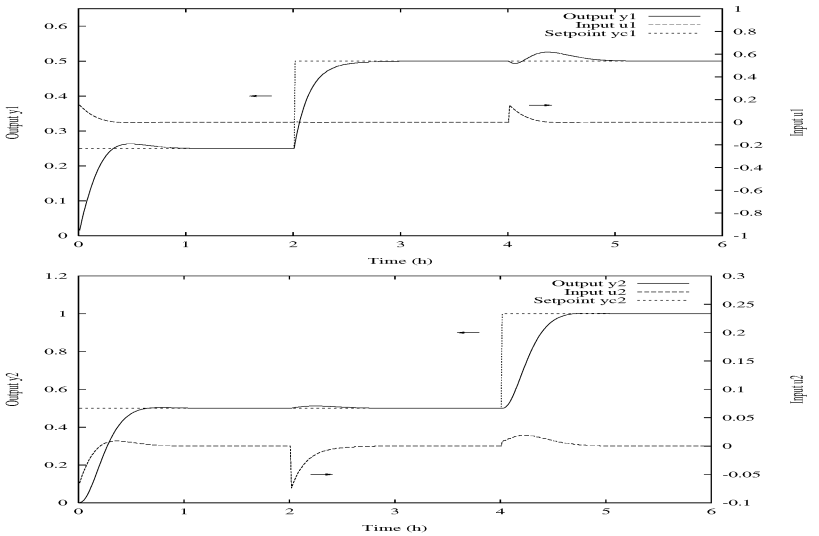


Figure 14.9: Linear quadratic control of Gilles extractive distillation column (criterion weighting: $\boldsymbol{Q} = 0.1\,\boldsymbol{I}$; $\boldsymbol{R} = \boldsymbol{I}$). Top: input u_1 and output y_1. Bottom: input u_2 and output y_2

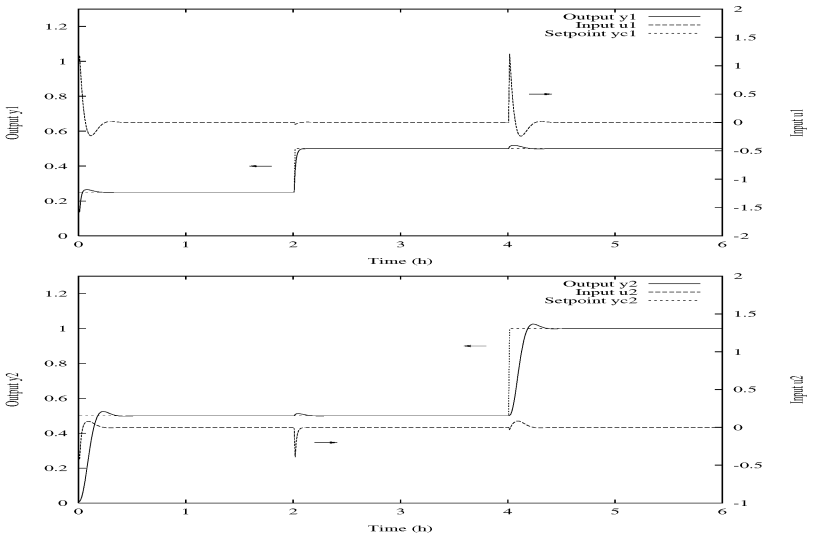


Figure 14.10: Linear quadratic control of Gilles extractive distillation column (criterion weighting: $\boldsymbol{Q} = \boldsymbol{I}$; $\boldsymbol{R} = 0.1\,\boldsymbol{I}$). Top: input u_1 and output y_1. Bottom: input u_2 and output y_2

14.6.2 Linear Quadratic Gaussian Control

In linear quadratic control, such as was previously discussed, the states are assumed to be perfectly known. In fact, this is seldom the case. Indeed, often the states have no physical reality and, if they have one, frequently they are not measurable or unmeasured. Thus, it is necessary to estimate the states to use their estimation in the control model.

In continuous time, the system is represented in the state space by the stochastic linear model

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{A}\,\boldsymbol{x}(t) + \boldsymbol{B}\,\boldsymbol{u}(t) + \boldsymbol{w}(t) \\ \boldsymbol{y}(t) = \boldsymbol{C}\,\boldsymbol{x}(t) + \boldsymbol{v}(t) \end{cases} \tag{14.260}$$

where $\boldsymbol{w}(t)$ and $\boldsymbol{v}(t)$ are uncorrelated Gaussian white noises, respectively of state and measurement (or output), of the respective covariance matrices

$$\mathrm{E}\left\{\boldsymbol{w}\boldsymbol{w}^T\right\} = \boldsymbol{W} \geq 0\,, \quad \mathrm{E}\left\{\boldsymbol{v}\boldsymbol{v}^T\right\} = \boldsymbol{V} > 0 \tag{14.261}$$

Denote by $\hat{\boldsymbol{x}}$ the state estimation, so that the state reconstruction error is $\boldsymbol{e}(t) = \boldsymbol{x} - \hat{\boldsymbol{x}}$. An optimal complete observer such as

$$\dot{\hat{\boldsymbol{x}}} = \boldsymbol{A}\,\hat{\boldsymbol{x}}(t) + \boldsymbol{B}\,\boldsymbol{u}(t) + \boldsymbol{K}_f(t)[y(t) - \boldsymbol{C}\,\hat{\boldsymbol{x}}(t)] \tag{14.262}$$

minimizes the covariance matrix of the state reconstruction error, thus

$$\mathrm{E}\left\{(\boldsymbol{x} - \hat{\boldsymbol{x}})\boldsymbol{P}_w(\boldsymbol{x} - \hat{\boldsymbol{x}})^T\right\} \tag{14.263}$$

where $\boldsymbol{P}_w$ is a weighting matrix (possibly, the identity matrix).

Kalman and Bucy (1961) solved this problem and showed that the estimator gain matrix $\boldsymbol{K}_f$ is equal to

$$\boldsymbol{K}_f(t) = \boldsymbol{P}_f(t)\,\boldsymbol{C}^T\,\boldsymbol{V}^{-1} \tag{14.264}$$

where $\boldsymbol{P}_f(t)$ is the solution of the continuous differential Riccati equation

$$\begin{aligned} \dot{\boldsymbol{P}}_f(t) \quad &= \boldsymbol{A}\,\boldsymbol{P}_f(t) + \boldsymbol{P}_f(t)\,\boldsymbol{A}^T - \boldsymbol{P}_f(t)\,\boldsymbol{C}^T\,\boldsymbol{V}^{-1}\,\boldsymbol{C}\,\boldsymbol{P}_f(t) + \boldsymbol{W} \\ &\text{with:} \quad \boldsymbol{P}_f(t_0) = \boldsymbol{P}_0 \end{aligned} \tag{14.265}$$

Moreover, the initial estimator condition is

$$\hat{\boldsymbol{x}}(t_0) = \hat{\boldsymbol{x}}_0 \tag{14.266}$$

The Kalman-Bucy filter thus calculated is the best state estimator or observer in the sense of linear least squares. It must be noticed that the determination of the Kalman filter is a dual problem of the linear quadratic optimal control problem: to go from the control problem to the estimation one, it suffices to make the following correspondences: $\boldsymbol{A} \to \boldsymbol{A}^T$, $\boldsymbol{B} \to \boldsymbol{C}^T$, $\boldsymbol{M}^T\boldsymbol{Q}\boldsymbol{M} \to \boldsymbol{W}$, $\boldsymbol{R} \to \boldsymbol{V}$, $\boldsymbol{P}_c \to \boldsymbol{P}_f$; on the one hand, the control Riccati equation progresses backwards with respect to time, on the other hand, the

estimation Riccati equation progresses forwards with respect to time. This latter remark obliges us to carefully manipulate all time-depending functions of the solutions of the Riccati equations (Kwakernaak and Sivan, 1972): $\boldsymbol{P}_c(t)$ (control problem) is equal to $\boldsymbol{P}_f(t_0 + t_f - t)$ (estimation problem), where t_0 is the initial time of the estimation problem and t_f the final time of the control problem.

When the estimation horizon becomes very large, in general the solution of the Riccati Eq. (14.265) tends towards a steady-state value, corresponding to the solution of the following algebraic Riccati equation

$$\boldsymbol{A}\,\boldsymbol{P}_f + \boldsymbol{P}_f\,\boldsymbol{A}^T - \boldsymbol{P}_f\,\boldsymbol{C}^T\,\boldsymbol{V}^{-1}\,\boldsymbol{C}\,\boldsymbol{P}_f + \boldsymbol{W} = 0 \tag{14.267}$$

giving the steady-state gain matrix of the estimator

$$\boldsymbol{K}_f = \boldsymbol{P}_f\,\boldsymbol{C}^T\,\boldsymbol{V}^{-1} \tag{14.268}$$

Kwakernaak and Sivan (1972) detail the conditions of convergence. For reasons of duality, the solving of the algebraic Riccati equation (14.267) is completely similar to that of Eq. (14.234).

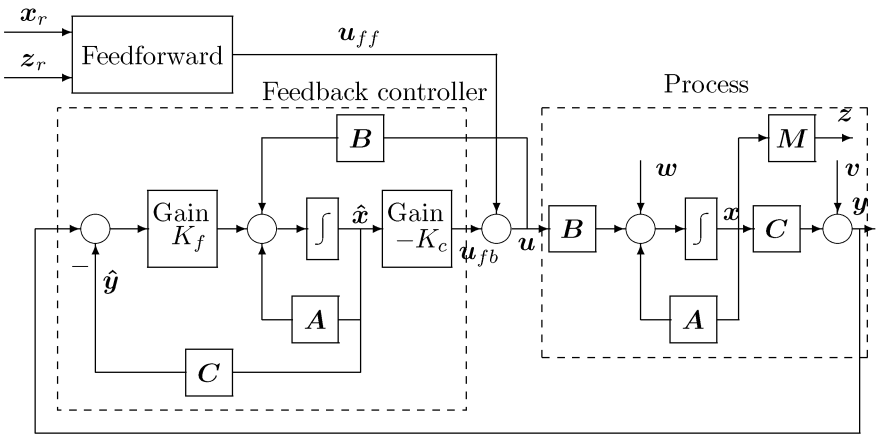


Figure 14.11: Structure of linear quadratic Gaussian control

Consider the general case of tracking. The control law similar to (14.243) is now based on the state estimation

$$\boldsymbol{u}^*(t) = -\boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{P}_c(t)\,\hat{\boldsymbol{x}}(t) + \boldsymbol{R}^{-1}\,\boldsymbol{B}^T\,\boldsymbol{s}(t) = -\boldsymbol{K}_c\hat{\boldsymbol{x}}(t) + \boldsymbol{u}_{ff}(t) \tag{14.269}$$

The system state equation can be written as

$$\dot{\hat{\boldsymbol{x}}}(t) = \boldsymbol{A}\,\boldsymbol{x}(t) - \boldsymbol{B}\,\boldsymbol{K}_c(t)\,\hat{\boldsymbol{x}}(t) + \boldsymbol{B}\,\boldsymbol{u}_{ff}(t) + \boldsymbol{w}(t) \tag{14.270}$$

so that the complete scheme of the Kalman filter and state feedback optimal control (Fig. 14.11) is written as

$$\begin{bmatrix} \dot{\boldsymbol{x}} \\ \dot{\hat{\boldsymbol{x}}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B}\,\boldsymbol{K}_c \\ \boldsymbol{K}_f\,\boldsymbol{C} & \boldsymbol{A} - \boldsymbol{K}_f\,\boldsymbol{C} - \boldsymbol{B}\,\boldsymbol{K}_c \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \hat{\boldsymbol{x}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{B}\,\boldsymbol{u}_{ff}(t) + \boldsymbol{w} \\ \boldsymbol{B}\,\boldsymbol{u}_{ff}(t) + \boldsymbol{K}_f\,\boldsymbol{v} \end{bmatrix} \tag{14.271}$$

which can be transformed by use of the estimation error $\boldsymbol{e}(t) = \boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)$

$$\begin{bmatrix} \dot{\boldsymbol{x}} \\ \dot{\boldsymbol{e}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{B}\,\boldsymbol{K}_c & \boldsymbol{B}\,\boldsymbol{K}_c \\ \boldsymbol{0} & \boldsymbol{A} - \boldsymbol{K}_f\,\boldsymbol{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{e} \end{bmatrix} + \begin{bmatrix} \boldsymbol{B}\,\boldsymbol{u}_{ff}(t) + \boldsymbol{w} \\ \boldsymbol{w} - \boldsymbol{K}_f\,\boldsymbol{v} \end{bmatrix} \quad (14.272)$$

The closed-loop eigenvalues are the union of the eigenvalues of the state feedback optimal control scheme and the eigenvalues of Kalman filter. Thus, it is possible to separately determine the observer and the state feedback optimal control law, which constitutes the separation principle of linear quadratic Gaussian control. This property that we have just verified for a complete observer is also verified for a reduced observer.

It is useful to notice that the Kalman filter gain is proportional to $\boldsymbol{P}$ (which will vary, but must be initialized) and inversely proportional to the measurement covariance matrix $\boldsymbol{V}$. Thus, if $\boldsymbol{V}$ is low, the filter gain will be very large, as the confidence in the measurement will be large; the risk of low robustness is then high. The Kalman filter can strongly deteriorate the stability margins (Doyle, 1978). The characteristic matrices of the Kalman filter can also be considered as tuning parameters. It is also possible to introduce an integrator per input-output channel in order to effectively realize the set point tracking; the modelled system represents, in this case, the group of the process plus the integrators. It is possible to add an output feedback (Lin, 1994) according to Fig. 14.12, which improves the robustness of regulation and tracking.

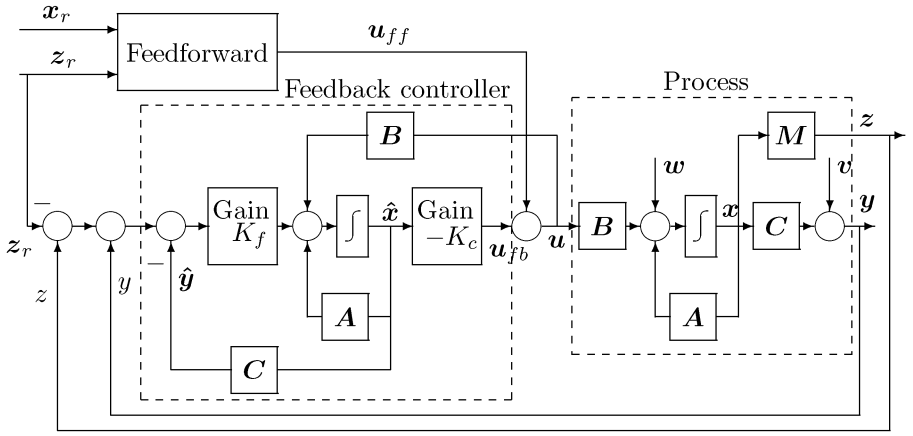


Figure 14.12: Structure of linear quadratic Gaussian control with added output feedback

An important development in LQG control is the taking into account of robustness so as to satisfy frequency criteria concerning the sensitivity and complementary sensitivity functions. Actually, the stability margins of LQG control may reveal themselves to be insufficient. LQG/LTR introducing loop transfer recovery (recovery of the properties at the process input) (Stein and Athans, 1987), (Maciejowski, 1989) allows us to solve this very important type of problem in state-space multivariable control.

Example 14.7: Linear Quadratic Gaussian Control of an Extractive Distillation Column

Again, we consider the control of the Gilles extractive distillation column, previously treated in the case where all the states are known. Now, suppose that the noise $\boldsymbol{w}(t)$ bearing on the state derivatives is Gaussian and has a standard deviation equal to 0.2, thus constituting a model error. The measurement noise is Gaussian and has a standard deviation equal to 0.5 . The initial states are all taken to be equal to 0.1 .

Consider the following values of the criterion weighting: $\boldsymbol{Q} = \boldsymbol{I}$; $\boldsymbol{R} = \boldsymbol{I}$ and the covariance matrices: $\boldsymbol{V} = 0.04\,\boldsymbol{I}$; $\boldsymbol{W} = 0.25\,\boldsymbol{I}$. Thus, the solution of the algebraic Riccati equation obtained from the Hamiltonian matrix method and concerning the Kalman estimator is

$$\boldsymbol{P}_f = \begin{bmatrix} 4.125 \times 10^{-3} & 1.339 \times 10^{-8} & -1.008 \times 10^{-9} & -3.999 \times 10^{-10} \\ 1.339 \times 10^{-8} & 1.912 \times 10^{-2} & -2.818 \times 10^{-3} & -7.681 \times 10^{-4} \\ -1.008 \times 10^{-9} & -2.818 \times 10^{-3} & 1.426 \times 10^{-2} & 1.315 \times 10^{-4} \\ -3.999 \times 10^{-10} & -7.681 \times 10^{-4} & 1.315 \times 10^{-4} & 4.034 \times 10^{-3} \end{bmatrix} \tag{14.273}$$

giving the steady-state gain matrix of the Kalman estimator

$$\boldsymbol{K}_f = \begin{bmatrix} 1.839 \times 10^{-7} & 2.499 \times 10^{-7} \\ 5.143 \times 10^{-1} & 4.801 \times 10^{-1} \\ -2.603 \times 10^{+0} & -8.218 \times 10^{-2} \\ -2.399 \times 10^{-2} & -2.521 \times 10^{+0} \end{bmatrix} \tag{14.274}$$

The values of the weighting matrices of the criterion and the covariance matrices are indicated in each case in the legends of Figs. 14.13 to 14.16. Figures 14.13 and 14.14 correspond to a state feedback control where no special account is taken of the measured output, while in Figs. 14.15 and 14.16, an output feedback acting on the reference, as indicated in Fig. 14.12, has been used. Naturally, the latter type of control gives better set point tracking.

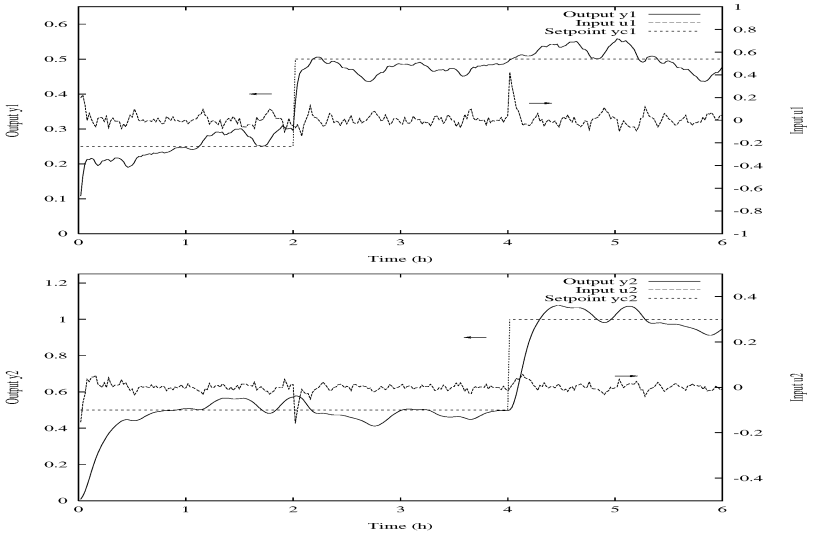


Figure 14.13: Linear quadratic Gaussian control of Gilles extractive distillation column (criterion weighting: $\boldsymbol{Q} = \boldsymbol{I}\,;\ \boldsymbol{R} = \boldsymbol{I}$. Initial $\boldsymbol{P}_f(0) = 0.1\,\boldsymbol{I}$. Covariance matrices: $\boldsymbol{V} = 0.04\,\boldsymbol{I}\,;\ \boldsymbol{W} = 0.25\,\boldsymbol{I}$). Top: input u_1 and output y_1. Bottom: input u_2 and output y_2

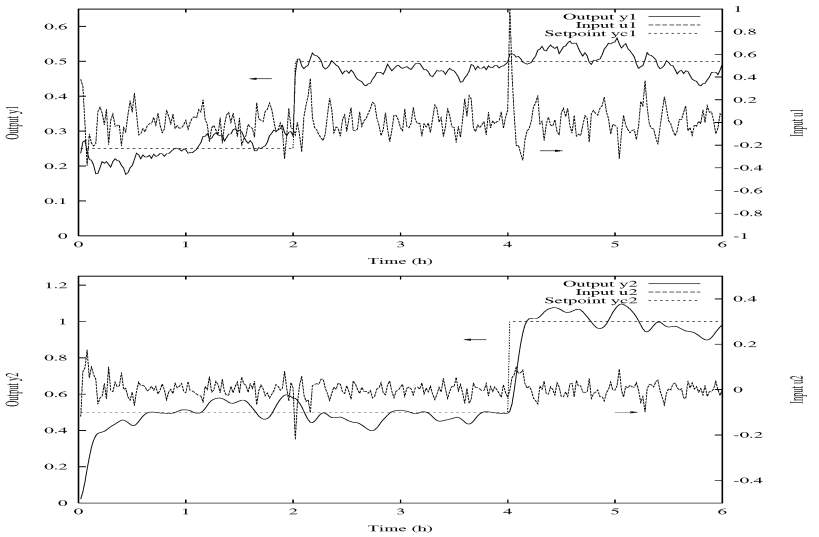


Figure 14.14: Linear quadratic Gaussian control of Gilles extractive distillation column (criterion weighting: $\boldsymbol{Q} = \boldsymbol{I}\,;\ \boldsymbol{R} = 0.1\,\boldsymbol{I}$. Initial $\boldsymbol{P}_f(0) = 0.1\,\boldsymbol{I}$. Covariance matrices: $\boldsymbol{V} = 0.04\,\boldsymbol{I}\,;\ \boldsymbol{W} = 0.25\,\boldsymbol{I}$). Top: input u_1 and output y_1. Bottom: input u_2 and output y_2

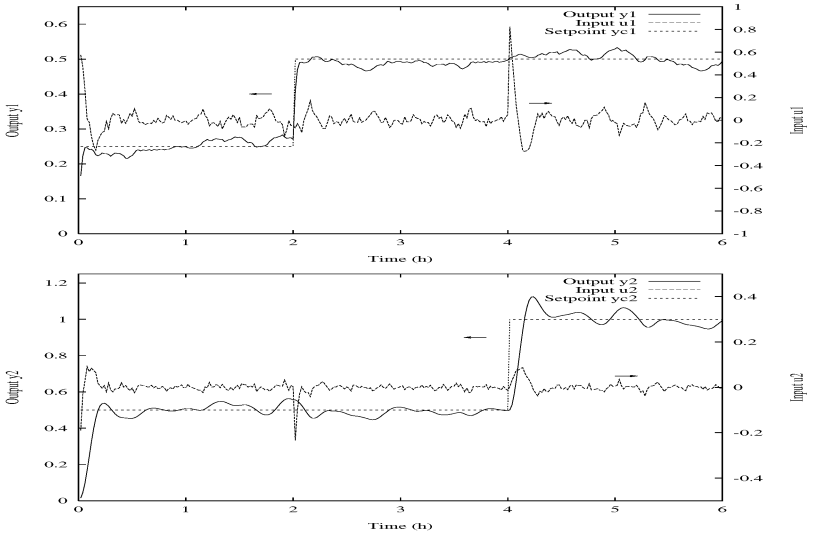


Figure 14.15: Linear quadratic Gaussian control of Gilles extractive distillation column, with output feedback (criterion weighting: $\boldsymbol{Q} = \boldsymbol{I}$; $\boldsymbol{R} = \boldsymbol{I}$. Initial $\boldsymbol{P}_f(0) = 0.1\,\boldsymbol{I}$. Covariance matrices: $\boldsymbol{V} = 0.04\,\boldsymbol{I}$; $\boldsymbol{W} = 0.25\,\boldsymbol{I}$). Top: input u_1 and output y_1. Bottom: input u_2 and output y_2

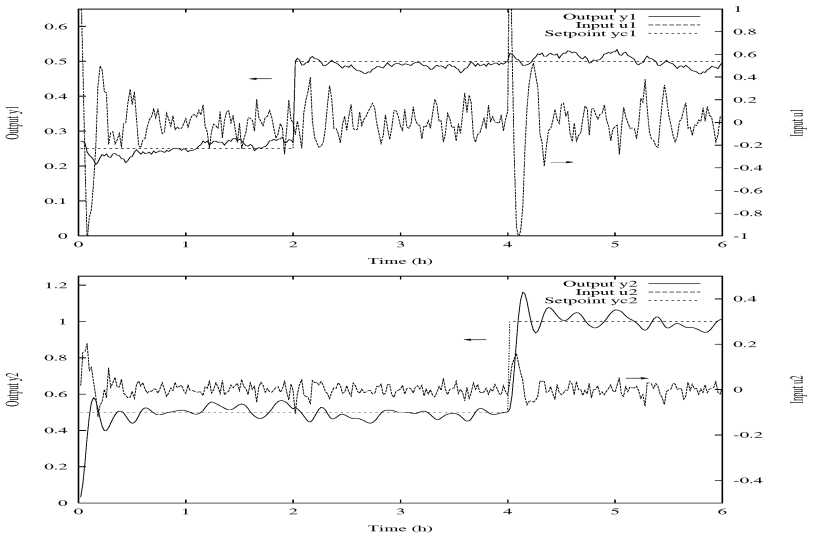


Figure 14.16: Linear quadratic Gaussian control of Gilles extractive distillation column, with output feedback (criterion weighting: $\boldsymbol{Q} = \boldsymbol{I}$; $\boldsymbol{R} = 0.1\,\boldsymbol{I}$. Initial $\boldsymbol{P}_f(0) = 0.1\,\boldsymbol{I}$. Covariance matrices: $\boldsymbol{V} = 0.04\,\boldsymbol{I}$; $\boldsymbol{W} = 0.25\,\boldsymbol{I}$). Top: input u_1 and output y_1. Bottom: input u_2 and output y_2

14.6.3 Discrete-Time Linear Quadratic Control

In discrete time, the system is represented in the state space by the deterministic linear model

$$\left\{ \begin{array}{ll} \boldsymbol{x}_{k+1} & = \boldsymbol{F}\,\boldsymbol{x}_k + \boldsymbol{G}\,\boldsymbol{u}_k \\ \boldsymbol{y}_k & = \boldsymbol{H}\,\boldsymbol{x}_k \end{array} \right. \tag{14.275}$$

The sought control must minimize a quadratic criterion similar to Eq. (14.216) thus

$$\begin{aligned} J \quad &= 0.5\,[\boldsymbol{z}_N^r - \boldsymbol{z}_N]^T\,\boldsymbol{Q}_N\,[\boldsymbol{z}_N^r - \boldsymbol{z}_N] \\ &+0.5\sum_{k=0}^{N-1}\left\{[\boldsymbol{z}_k^r - \boldsymbol{z}_k]^T\,\boldsymbol{Q}\,[\boldsymbol{z}_k^r - \boldsymbol{z}_k] + \boldsymbol{u}_k^T\,\boldsymbol{R}\,\boldsymbol{u}_k\right\} \\ &= 0.5\,[\boldsymbol{x}_N^r - \boldsymbol{x}_N]^T\,\boldsymbol{M}^T\,\boldsymbol{Q}_N\,\boldsymbol{M}\,[\boldsymbol{x}_N^r - \boldsymbol{x}_N] \\ &+0.5\sum_{k=0}^{N-1}\left\{[\boldsymbol{x}_k^r - \boldsymbol{x}_k]^T\,\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M}\,[\boldsymbol{x}_k^r - \boldsymbol{x}_k] + \boldsymbol{u}_k^T\,\boldsymbol{R}\,\boldsymbol{u}_k\right\} \end{aligned} \tag{14.276}$$

where the matrices $\boldsymbol{Q}_N$, $\boldsymbol{Q}$ are semipositive definite and $\boldsymbol{R}$ is positive definite. Furthermore $\boldsymbol{z} = \boldsymbol{M}\boldsymbol{x}$ represents measurements or outputs. It would be possible to use variational methods to deduce from them the optimal control law (Borne et al., 1990; Lewis, 1986), which would provide a system perfectly similar to Eq. (14.223). However, variational methods are a priori designed in the framework of continuous variables, thus for continuous time; on the other hand, dynamic programming is perfectly adapted to the discrete case. Thus, we will sketch out the reasoning in this framework. For more details, it is possible, for example, to refer to Dorato and Levis (1971); Foulard et al. (1987).

The system is considered at any instant i included in the interval $[0, N]$, assuming that the policy preceding that instant, thus the sequence of the $\{\boldsymbol{u}_k, k \in [i+1, N]\}$, is optimal (the final instant N is the starting point for performing the procedure of dynamic programming). In these conditions, the criterion of interest is in the form

$$\begin{aligned} J_i \quad &= 0.5\,[\boldsymbol{z}_N^r - \boldsymbol{z}_N]^T\,\boldsymbol{Q}_N\,[\boldsymbol{z}_N^r - \boldsymbol{z}_N] \\ &+0.5\sum_{k=i}^{N-1}\left\{[\boldsymbol{z}_k^r - \boldsymbol{z}_k]^T\,\boldsymbol{Q}\,[\boldsymbol{z}_k^r - \boldsymbol{z}_k] + \boldsymbol{u}_k^T\,\boldsymbol{R}\,\boldsymbol{u}_k\right\} \\ &= 0.5\,[\boldsymbol{x}_N^r - \boldsymbol{x}_N]^T\,\boldsymbol{M}^T\,\boldsymbol{Q}_N\,\boldsymbol{M}\,[\boldsymbol{x}_N^r - \boldsymbol{x}_N] + \sum_{k=i}^{N-1} L_k(\boldsymbol{x}_k, \boldsymbol{u}_k) \end{aligned} \tag{14.277}$$

with the revenue function L_k defined by

$$L_k(\boldsymbol{x}_k, \boldsymbol{u}_k) = 0.5\left\{[\boldsymbol{x}_k^r - \boldsymbol{x}_k]^T\,\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M}\,[\boldsymbol{x}_k^r - \boldsymbol{x}_k] + \boldsymbol{u}_k^T\,\boldsymbol{R}\,\boldsymbol{u}_k\right\} \tag{14.278}$$

According to the Bellman optimality principle, the optimal value J_i^* of the criterion can be expressed in a recurrent form

$$J_i^* = \min_{\boldsymbol{u}}\left\{0.5[\boldsymbol{x}_i^r - \boldsymbol{x}_i]^T\,\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M}\,[\boldsymbol{x}_i^r - \boldsymbol{x}_i] + \boldsymbol{u}_i^T\,\boldsymbol{R}\,\boldsymbol{u}_i + J_{i+1}^*\right\} \tag{14.279}$$

The final state is assumed to be free. It is necessary to know the expression of J^*_{i+1} with respect to the state to be able to perform the minimization. Reasoning by recurrence, suppose that it is in quadratic form

$$J^*_{i+1} = 0.5 \left\{ \boldsymbol{x}^T_{i+1} \boldsymbol{S}_{i+1} \boldsymbol{x}_{i+1} + 2\, \boldsymbol{g}_{i+1} \boldsymbol{x}_{i+1} + h_{i+1} \right\} \tag{14.280}$$

hence

$$\begin{aligned} J^*_{i+1} = \quad & 0.5 \left\{ [\boldsymbol{F} \boldsymbol{x}_i + \boldsymbol{G} \boldsymbol{u}_i]^T \boldsymbol{S}_{i+1} [\boldsymbol{F} \boldsymbol{x}_i + \boldsymbol{G} \boldsymbol{u}_i] \right. \\ & \left. + 2\, \boldsymbol{g}_{i+1} [\boldsymbol{F} \boldsymbol{x}_i + \boldsymbol{G} \boldsymbol{u}_i] + h_{i+1} \right\} \end{aligned} \tag{14.281}$$

We deduce

$$\begin{aligned} J^*_i = \quad & 0.5 \min_{\boldsymbol{u}_i} \left\{ [\boldsymbol{x}^r_i - \boldsymbol{x}_i]^T \boldsymbol{M}^T \boldsymbol{Q} \boldsymbol{M} [\boldsymbol{x}^r_i - \boldsymbol{x}_i] + \boldsymbol{u}^T_i \boldsymbol{R} \boldsymbol{u}_i \right. \\ & + [\boldsymbol{F} \boldsymbol{x}_i + \boldsymbol{G} \boldsymbol{u}_i]^T \boldsymbol{S}_{i+1} [\boldsymbol{F} \boldsymbol{x}_i + \boldsymbol{G} \boldsymbol{u}_i] \\ & \left. + 2\, \boldsymbol{g}_{i+1} [\boldsymbol{F} \boldsymbol{x}_i + \boldsymbol{G} \boldsymbol{u}_i] + h_{i+1} \right\} \end{aligned} \tag{14.282}$$

We search the minimum with respect to $\boldsymbol{u}_i$ thus

$$\boldsymbol{R} \boldsymbol{u}^*_i + \boldsymbol{G}^T \boldsymbol{S}_{i+1} [\boldsymbol{F} \boldsymbol{x}_i + \boldsymbol{G} \boldsymbol{u}^*_i] + \boldsymbol{G}^T \boldsymbol{g}_{i+1} = 0 \tag{14.283}$$

or

$$\boldsymbol{u}^*_i = -[\boldsymbol{R} + \boldsymbol{G}^T \boldsymbol{S}_{i+1} \boldsymbol{G}]^{-1} [\boldsymbol{G}^T \boldsymbol{S}_{i+1} \boldsymbol{F} \boldsymbol{x}_i + \boldsymbol{G}^T \boldsymbol{g}_{i+1}] \tag{14.284}$$

provided that the matrix $[\boldsymbol{R} + \boldsymbol{G}^T \boldsymbol{S}_{i+1} \boldsymbol{G}]$ is invertible. Notice that the optimal control is in the form

$$\boldsymbol{u}^*_i = -\boldsymbol{K}_i \boldsymbol{x}_i + \boldsymbol{k}_i \boldsymbol{g}_{i+1} \tag{14.285}$$

revealing the state feedback with the gain matrix $\boldsymbol{K}_i$ and feedforward with the gain $\boldsymbol{k}_i$. Thus, we set

$$\begin{aligned} \boldsymbol{K}_i \quad &= [\boldsymbol{R} + \boldsymbol{G}^T \boldsymbol{S}_{i+1} \boldsymbol{G}]^{-1} \boldsymbol{G}^T \boldsymbol{S}_{i+1} \boldsymbol{F} \; ; \qquad \boldsymbol{S}_N = \boldsymbol{M}^T \boldsymbol{Q}_N \boldsymbol{M} \\ \boldsymbol{k}_i \quad &= -[\boldsymbol{R} + \boldsymbol{G}^T \boldsymbol{S}_{i+1} \boldsymbol{G}]^{-1} \boldsymbol{G}^T \end{aligned} \tag{14.286}$$

It is then possible to verify that J^*_i is effectively in quadratic form; thus we find

$$\begin{aligned} \boldsymbol{S}_i \quad &= \boldsymbol{M}^T \boldsymbol{Q} \boldsymbol{M} + \boldsymbol{F}^T \boldsymbol{S}_{i+1} (\boldsymbol{F} - \boldsymbol{G} \boldsymbol{K}_i) \\ \boldsymbol{g}_i \quad &= -\boldsymbol{M}^T \boldsymbol{Q} \boldsymbol{z}^r_i + (\boldsymbol{F}^T - \boldsymbol{G} \boldsymbol{K}_i)^T \boldsymbol{g}_{i+1} \; ; \qquad \boldsymbol{g}_N = \boldsymbol{M} \boldsymbol{Q}_N \boldsymbol{z}^r_N \end{aligned} \tag{14.287}$$

The group of Eqs. (14.285), (14.286), (14.287) allows us to determine the inputs $\boldsymbol{u}$. When not all states are known, of course it is necessary to use a discrete Kalman filter which will work with the optimal control law according to the same separation principle as in the continuous case.

It can be shown that Eq. (14.287) is equivalent to the discrete differential Riccati equation

$$\boldsymbol{S}_i = (\boldsymbol{F} - \boldsymbol{G} \boldsymbol{K}_i)^T \boldsymbol{S}_{i+1} (\boldsymbol{F} - \boldsymbol{G} \boldsymbol{K}_i) + \boldsymbol{K}^T_i \boldsymbol{R} \boldsymbol{K}_i + \boldsymbol{M}^T \boldsymbol{Q} \boldsymbol{M} \tag{14.288}$$

presented in Joseph form and better adapted to numerical calculation.

Let us apply the Hamilton-Jacobi principle to the discrete-time optimal regulator. If the control law $\boldsymbol{u}_i^*$ and the corresponding states $\boldsymbol{x}_i^*$ are optimal, according to the Hamilton-Jacobi principle, there exists a costate vector $\boldsymbol{\psi}_i^*$ such that $\boldsymbol{u}_i^*$ is the value of the control $\boldsymbol{u}_i$ which maximizes the Hamiltonian function H_a

$$\begin{aligned} H_a &= -L_i(\boldsymbol{x}_i^*, \boldsymbol{u}_i) + \boldsymbol{\psi}_{i+1}^{*T}\, \boldsymbol{x}_{i+1} \\ &= -L_i(\boldsymbol{x}_i^*, \boldsymbol{u}_i) + \boldsymbol{\psi}_{i+1}^{*T}\, [\boldsymbol{F}\, \boldsymbol{x}_i + \boldsymbol{G}\, \boldsymbol{u}_i] \\ &= -0.5 \left\{ [\boldsymbol{x}_i^r - \boldsymbol{x}_i]^T\, \boldsymbol{M}^T\, \boldsymbol{Q}\, \boldsymbol{M}\, [\boldsymbol{x}_i^r - \boldsymbol{x}_i] + \boldsymbol{u}_i^T\, \boldsymbol{R}\, \boldsymbol{u}_i \right\} + \boldsymbol{\psi}_{i+1}^{*T}\, [\boldsymbol{F}\, \boldsymbol{x}_i + \boldsymbol{G}\, \boldsymbol{u}_i] \end{aligned} \tag{14.289}$$

The Hamilton-Jacobi conditions give

$$\boldsymbol{\psi}_i^* = -\frac{\partial H_a}{\partial \boldsymbol{x}_i^*} = \boldsymbol{M}^T\, \boldsymbol{Q}\, \boldsymbol{M}\, [\boldsymbol{x}_i^* - \boldsymbol{x}_i^r] - \boldsymbol{F}^T\, \boldsymbol{\psi}_{i+1} \quad \text{with: } \boldsymbol{\psi}_N = \boldsymbol{M}^T\, \boldsymbol{Q}\, \boldsymbol{M}\, [\boldsymbol{x}_N^* - \boldsymbol{x}_N^r] \tag{14.290}$$

and the control which maximizes the Hamiltonian function H_a is such that

$$\frac{dH_a}{du_i} = 0 \Longrightarrow -\boldsymbol{R}\, \boldsymbol{u}_i + \boldsymbol{G}^T\, \boldsymbol{\psi}_{i+1} = 0 \Longrightarrow \boldsymbol{u}_i^* = \boldsymbol{R}^{-1}\, \boldsymbol{G}^T\, \boldsymbol{\psi}_{i+1} \tag{14.291}$$

hence

$$\boldsymbol{x}_{i+1} = \boldsymbol{F}\, \boldsymbol{x}_i + \boldsymbol{G}\, \boldsymbol{u}_i = \boldsymbol{F}\, \boldsymbol{x}_i + \boldsymbol{G}\, \boldsymbol{R}^{-1}\, \boldsymbol{G}^T\, \boldsymbol{\psi}_{i+1} \tag{14.292}$$

Introduce the matrix $\mathcal{H}$ such that

$$\begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{\psi}_i \end{bmatrix} = \mathcal{H} \begin{bmatrix} \boldsymbol{x}_{i+1} \\ \boldsymbol{\psi}_{i+1} \end{bmatrix} \qquad \text{or :} \qquad \begin{bmatrix} \boldsymbol{x}_{i+1} \\ \boldsymbol{\psi}_{i+1} \end{bmatrix} = \mathcal{H}^{-1} \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{\psi}_i \end{bmatrix} \tag{14.293}$$

Assume $\boldsymbol{x}_i^r = 0$ for the regulation case. The two conditions (14.290) and (14.291) can be grouped as

$$\begin{aligned} \boldsymbol{\psi}_i^* &= \boldsymbol{M}^T\, \boldsymbol{Q}\, \boldsymbol{M}\, \boldsymbol{x}_i^* - \boldsymbol{F}^T\, \boldsymbol{\psi}_{i+1} \\ \boldsymbol{x}_{i+1} &= \boldsymbol{F}\, \boldsymbol{x}_i + \boldsymbol{G}\, \boldsymbol{R}^{-1}\, \boldsymbol{G}^T\, \boldsymbol{\psi}_{i+1} \end{aligned} \tag{14.294}$$

from which we deduce the matrix $\mathcal{H}$

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \boldsymbol{F}^{-1} & -\boldsymbol{F}^{-1}\, \boldsymbol{G}\, \boldsymbol{R}^{-1}\, \boldsymbol{G}^T \\ \boldsymbol{M}^T\, \boldsymbol{Q}\, \boldsymbol{M}\, \boldsymbol{F}^{-1} & -\boldsymbol{F}^T - \boldsymbol{M}^T\, \boldsymbol{Q}\, \boldsymbol{M}\, \boldsymbol{F}^{-1}\, \boldsymbol{G}\, \boldsymbol{R}^{-1}\, \boldsymbol{G}^T \end{bmatrix} \tag{14.295}$$

In the case where a steady-state gain $\boldsymbol{K}_\infty$ is satisfactory, which can be realized when the horizon N is large, the gain matrix can be obtained after solving the algebraic Riccati equation

$$\boldsymbol{S} = \boldsymbol{F}^T\, [\boldsymbol{S} - \boldsymbol{S}\, \boldsymbol{G}\, (\boldsymbol{G}^T\, \boldsymbol{S}\, \boldsymbol{G} + \boldsymbol{R})^{-1}\, \boldsymbol{G}^T\, \boldsymbol{S}]\, \boldsymbol{F} + \boldsymbol{M}\, \boldsymbol{Q}\, \boldsymbol{M} \tag{14.296}$$

whose solution (corresponding to discrete time) is obtained in a parallel manner to the continuous case, by first considering the matrix $\boldsymbol{\mathcal{H}}$. Its inverse is the

symplectic[7] matrix $\boldsymbol{\mathcal{H}}^{-1}$ equal to

$$\boldsymbol{\mathcal{H}}^{-1} = \begin{bmatrix} \boldsymbol{F} + \boldsymbol{G}\,\boldsymbol{R}^{-1}\,\boldsymbol{G}^T\,\boldsymbol{F}^{-T}\,\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M} & -\boldsymbol{G}\,\boldsymbol{R}^{-1}\,\boldsymbol{G}^T\,\boldsymbol{F}^{-T} \\ \boldsymbol{F}^{-T}\,\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{M} & -\boldsymbol{F}^{-T} \end{bmatrix} \tag{14.297}$$

We seek the eigenvalues and associated eigenvectors of $\boldsymbol{\mathcal{H}}^{-1}$. Then, we form the matrix $\boldsymbol{U}$ of the eigenvectors so that the first n columns correspond to the stable eigenvalues (inside the unit circle), in the form

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} \\ \boldsymbol{U}_{21} & \boldsymbol{U}_{22} \end{bmatrix} \tag{14.298}$$

where the blocks $\boldsymbol{U}_{ij}$ have dimension $n \times n$. The solution of the discrete Riccati algebraic equation (14.296) is then

$$\boldsymbol{S}_\infty = \boldsymbol{U}_{21}\,\boldsymbol{U}_{11}^{-1} \tag{14.299}$$

giving the steady-state matrix.

For the tracking problem, in parallel to the stationary solution for the gain matrix $\boldsymbol{K}$, the stationary solution for the feedforward gain is deduced from Eq. (14.287) and is given by

$$\boldsymbol{g}_i = [\boldsymbol{I} - (\boldsymbol{F}^T - \boldsymbol{G}\,\boldsymbol{K}_\infty)^T]^{-1}\,[-\boldsymbol{M}^T\,\boldsymbol{Q}\,\boldsymbol{z}_i^r] \tag{14.300}$$

The use of stationary gains provides a sub-optimal solution but is more robust than using the optimal gains coming from Eqs. (14.286) and (14.287).

The discrete linear quadratic Gaussian control is derived from the previously described discrete linear quadratic control by coupling a discrete linear Kalman filter in order to estimate the states.

Remark

The operating conditions of quadratic control have not been discussed. In general, it is assumed that the pair $(\boldsymbol{A}, \boldsymbol{B})$ in continuous time, or $(\boldsymbol{F}, \boldsymbol{G})$ in discrete time, is controllable. Moreover, when the horizon is infinite and when we are looking for steady-state solutions of the Riccati equation, the condition that the pair $(\boldsymbol{A}, \sqrt{\boldsymbol{Q}})$ in continuous time, or $(\boldsymbol{F}, \sqrt{\boldsymbol{Q}})$ in discrete time, is observable (the notation $\boldsymbol{H} = \sqrt{\boldsymbol{Q}}$ means that $\boldsymbol{Q} = \boldsymbol{H}^T\,\boldsymbol{H}$) must be added.

References

B.D.O. Anderson and J.B. Moore. *Linear Optimal Control.* Prentice Hall, Englewood Cliffs, New Jersey, 1971.

[7] A matrix $\boldsymbol{A}$ is symplectic, when, given the matrix $J = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{0} \end{bmatrix}$, the matrix $\boldsymbol{A}$ verifies $\boldsymbol{A}^T\,\boldsymbol{J}\,\boldsymbol{A} = \boldsymbol{J}$.

If λ is an eigenvalue of a symplectic matrix $\boldsymbol{A}$, $1/\lambda$ is also an eigenvalue of $\boldsymbol{A}$; λ is thus also an eigenvalue of $\boldsymbol{A}^{-1}$ (Laub, 1979).

B.D.O. Anderson and J.B. Moore. *Optimal Control, Linear Quadratic Methods.* Prentice Hall, Englewood Cliffs, New Jersey, 1990.

R. Aris. Studies in optimization. II. Optimal temperature gradients in tubular reactors. *Chem. Eng. Sci.*, 13(1):18–29, 1960.

R. Aris. *The Optimal Design of Chemical Reactors: A Study in Dynamic Programming.* Academic Press, New York, 1961.

R. Aris, D.F. Rudd, and N.R. Amundson. On optimum cross current extraction. *Chem. Eng. Sci.*, 12:88–97, 1960.

W.F. Arnold and A.J. Laub. Generalized eigenproblem algorithms and software for algebraic Riccati equations. *IEEE Proceedings*, 72(12):1746–1754, 1984.

M. Athans and P.L. Falb. *Optimal Control: An Introduction to the Theory and its Applications.* Mac Graw Hill, New York, 1966.

J.R. Banga and E.F. Carrasco. Rebuttal to the comments of Rein Luus on "Dynamic optimization of batch reactors using adaptive stochastic algorithms". *Ind. Eng. Chem. res.*, 37:306–307, 1998.

R. Bellman. *Dynamic PProgramming.* Princeton University Press, Princeton, New Jersey, 1957.

R. Bellman and S. Dreyfus. *Applied Dynamic Programming.* Princeton University Press, Princeton, New Jersey, 1962.

L.T. Biegler. Solution of dynamic optimization problems by successive quadratic programming and orthogonal collocation. *Comp. Chem. Eng.*, 8:243–248, 1984.

B. Bojkov and R. Luus. Optimal control of nonlinear systems with unspecified final times. *Chem. Eng. Sci.*, 51(6):905–919, 1996.

P. Borne, G. Dauphin-Tanguy, J.P. Richard, F. Rotella, and I. Zambettakis. *Commande et Optimisation des Processus.* Technip, Paris, 1990.

R. Boudarel, J. Delmas, and P. Guichet. *Commande Optimale des Processus.* Dunod, Paris, 1969.

A.E. Bryson. *Dynamic Optimization.* Addison Wesley, Menlo Park, California, 1999.

A.E. Bryson and Y.C. Ho. *Applied Optimal Control.* Hemisphere, Washington, 1975.

E.F. Carrasco and J.R. Banga. Dynamic optimization of batch reactors using adaptive stochastic algorithms. *Ind. Eng. Chem. Res.*, 36:2252–2261, 1997.

H. Cartan. *Cours de Calcul Différentiel.* Hermann, Paris, 1967.

J.E. Cuthrell and L.T. Biegler. On the optimization of differential-algebraic process systems. *A.I.Ch.E. J.*, 33:1257–1270, 1987.

J. Dorato and A.H. Levis. *IEEE Trans. A. C.*, AC-16(6):613–620, 1971.

J.C. Doyle. Guaranteed margins for LQG regulators. *IEEE Trans. Automat. Control*, AC-23:756–757, 1978.

J.N. Farber and R.L. Laurence. The minimum time problem in batch radical polymerization: a comparison of two policies. *Chem. Eng. Commun.*, 46: 347–364, 1986.

A. Feldbaum. *Principes Théoriques des Systèmes Asservis Optimaux.* Mir, Moscou, 1973. Edition Française.

M. Fikar, M.A. Latifi, J.P. Corriou, and Y. Creff. Cvp-based optimal control of an industrial depropanizer column. *Comp. Chem. Engn.*, 24:909–915, 2000.

R. Fletcher. *Practical Methods of Optimization.* Wiley, Chichester, 1991.

C. Foulard, S. Gentil, and J.P. Sandraz. *Commande et Régulation par Calculateur Numérique.* Eyrolles, Paris, 1987.

C. Gentric, F. Pla, M.A. Latifi, and J.P. Corriou. Optimization and non-linear control of a batch emulsion polymerization reactor. *Chem. Eng. J.*, 75:31–46, 1999.

E.D. Gilles and B. Retzbach. Modeling, simulation and control of distillation columns with sharp temperature profiles. *IEEE Trans. Automat. Control*, AC-28(5):628–630, 1983.

E.D. Gilles, B. Retzbach, and F. Silberberger. Modeling, simulation and control of an extractive distillation column. In *Computer Applications to Chemical Engineering*, volume 124 of *ACS Symposium Series*, pages 481–492, 1980.

C.J. Goh and K.L. Teo. Control parametrization: a unified approach to optimal control problems with general constraints. *Automatica*, 24:3–18, 1988.

M.J. Grimble and M.A. Johnson. *Optimal Control and Stochastic Estimation: Deterministic Systems*, volume 1. Wiley, Chichester, 1988a.

M.J. Grimble and M.A. Johnson. *Optimal Control and Stochastic Estimation: Stochastic Systems*, volume 2. Wiley, Chichester, 1988b.

T. Kailath. *Linear Systems Theory.* Prentice Hall, Englewood Cliffs, New Jersey, 1980.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D, J. Basic Eng.*, 82:35–45, 1960.

R.E. Kalman. Mathematical description of linear dynamical systems. *J. SIAM Control*, series A:152–192, 1963.

R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME Ser. D, J. Basic Eng.*, 83:95–108, 1961.

A. Kaufmann and R. Cruon. *La Programmation Dynamique. Gestion Scientifique Séquentielle.* Dunod, Paris, 1965.

D.E. Kirk. *Optimal Control Theory. An Introduction.* Prentice Hall, Englewood Cliffs, New Jersey, 1970.

H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems.* Wiley-Interscience, New York, 1972.

Y.D. Kwon and L.B. Evans. A coordinate transformation method for the numerical solution of non-linear minimum-time control problems. *AIChE J.*, 21:1158–, 1975.

F. Lamnabhi-Lagarrigue. Singular optimal control problems: on the order of a singular arc. *Systems & control letters*, 9:173–182, 1987.

M.A. Latifi, J.P. Corriou, and M. Fikar. Dynamic optimization of chemical processes. *Trends in Chem. Eng.*, 4:189–201, 1998.

A.J. Laub. A Schur method for solving algebraic Riccati equations. *IEEE Trans. Automat. Control*, AC-24(6):913–921, 1979.

E.B. Lee and L. Markus. *Foundations of Optimal Control Theory.* Krieger, Malabar, Florida, 1967.

F.L. Lewis. *Optimal Control.* Wiley, New York, 1986.

C.F. Lin. *Advanced Control Systems Design.* Prentice Hall, Englewood Cliffs, New Jersey, 1994.

R. Luss. Application of dynamic programming to high-dimensional nonlinear optimal control systems. *Int. J. Cont.*, 52(1):239–250, 1990.

R. Luss. Application of iterative dynamic to very high-dimensional systems. *Hung. J. Ind. Chem.*, 21:243–250, 1993.

R. Luss and B. Bojkov. Application of iterative dynamic to time-optimal control. *Trans. Inst. Chem. Chem.*, 72:72–80, 1994.

R. Luus. Numerical convergence properties of iterative dynamic programming when applied to high dimensional systems. *Trans. IChemE, part A*, 74:55–62, 1996.

R. Luus and D. Hennessy. Optimization of fed-batch reactors by the Luus-Jaakola optimization procedure. *Ind. Eng. Chem. Res.*, 38:1948–1955, 1999.

J.M. Maciejowski. *Multivariable Feedback Design.* Addison-Wesley, Wokingham, England, 1989.

W. Mekarapiruk and R. Luus. Optimal control of inequality state constrained systems. *Ind. Eng. Chem. Res.*, 36:1686–1694, 1997.

L. Pontryaguine, V. Boltianski, R. Gamkrelidze, and E. Michtchenko. *Théorie Mathématique des Processus Optimaux.* Mir, Moscou, 1974. Edition Française.

L. Pun. *Introduction à la Pratique de l'Optimisation.* Dunod, Paris, 1972.

W.H. Ray and J. Szekely. *Process Optimization with Applications in Metallurgy and Chemical Engineering.* Wiley, New York, 1973.

S.M. Roberts. *Dynamic Programming in Chemical Engineering and Process Control.* Academic Press, New York, 1964.

S.M. Roberts and C.G. Laspe. Computer control of a thermal cracking reaction. *Ind. Eng. Chem.*, 53(5):343–348, 1961.

K. Schittkowski. NLPQL: A Fortran subroutine solving constrained nonlinear programming problems. *Ann. Oper. Res.*, 5:485–500, 1985.

R. Soeterboek. *Predictive Control - A Unified Approach.* Prentice Hall, Englewood Cliffs, New Jersey, 1992.

G. Stein and M. Athans. The LQG/LTR procedure for multivariable feedback control design. *IEEE Trans. Automat. Control*, AC-32(2):105–114, 1987.

K.L. Teo, C.J. Goh, and K.H. Wong. *A Unified Computational Approach to Optimal Control Problems.* Longman Scientific & Technical, Harlow, Essex, England, 1991.

Chapter 15

Generalized Predictive Control

The principle of predictive control makes it attractive for many applications, either as linear or nonlinear control. In this chapter, only linear generalized predictive control is studied. Model predictive control is treated in Chap. 16 where linear model predictive control is discussed in two forms and nonlinear predictive control is also discussed. The latter (Rawlings et al., 1994) is often treated as an optimization problem with constraints on the state and on the input, where we seek a control minimizing a technical-economic criterion.

15.1 Interest in Generalized Predictive Control

Generalized predictive control (GPC) proposed by Clarke et al. (1987a,b) is among the control methods that are usable for adaptive control. To be used, it must be coupled with an identification method of the process model; if the identification is realized on-line, it is adaptive control. Bitmead et al. (1990) devoted a entire book to GPC and, in particular, studied the interaction between parametric identification and the choice of control scheme, insisting on the interest in combining them for the robustness of the system rather than exciting oneself in a very deep aspect either of identification or of control, even more so as identification will have to be realized in closed loop. Bitmead et al. (1990) even qualify of synergy the relation between recursive least-squares identification and linear quadratic control, i.e. the robustness of the combination is better than that obtained by applying them separately. Also, the book by Camacho and Bordons (1998) deals mainly with GPC. It also presents multivariable GPC and includes several examples. On the other hand, this method has been used successfully in industrial applications in different forms (Clarke, 1988; Defaye et al., 1993; LeLann et al., 1986; Rafilamanana et al., 1992). Among the claimed advantages of GPC, Clarke et al. (1987a) mention that it can be applied to processes presenting a variable time delay, to nonminimum-phase

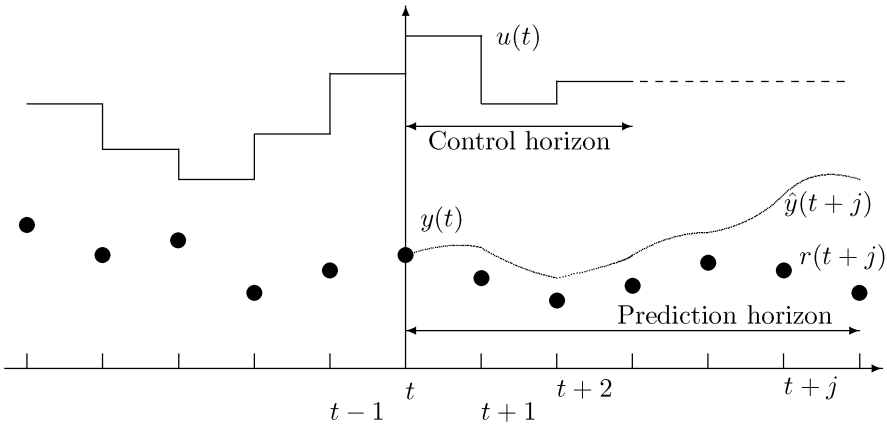


Figure 15.1: Principle of extended horizon predictive control

processes, and that it poses no apparent problem when the process model includes too many parameters, contrary to pole-placement or linear quadratic control.

Predictive control owes its name to the fact that it takes into account predictions of the process future outputs and sometimes also of the future inputs. The outputs and the inputs are predicted over a finite horizon.

15.2 Brief Overview of Predictive Control Evolution

Predictive control did not appear suddenly, but rather appeared as an evolution through the minimum-variance controller (Aström, 1970) and the self-tuning controller (Aström and Wittenmark, 1973) obtained by minimization of the criterion

$$J(u,t) = \mathrm{E}\left\{[y(t+d+1) - r(t+d+1)]^2\right\} \tag{15.1}$$

where r is the reference signal to be tracked by the output and d the delay of the output with respect to the input (excluding the minimum unit time delay of the output with respect to the input). Minimization of the criterion gives the input $u(t)$ to apply. The controller thus obtained is only convenient for minimum-phase systems. In order to extend it to nonminimum-phase systems, generalized minimum-variance control was proposed by Clarke and Gawthrop (1975, 1979), by introducing a penalty on the input, thus giving the new criterion

$$J(u,t) = \mathrm{E}\left\{[y(t+d+1) - r(t+d+1)]^2 + \lambda u(t)^2\right\} \tag{15.2}$$

or furthermore

$$J(u,t) = \mathrm{E}\left\{[y(t+d+1) - r(t+d+1)]^2 + \lambda \Delta u(t)^2\right\} \tag{15.3}$$

by introducing the input variation $\Delta u(t)$ to ensure a zero steady-state error in the case of a constant nonzero reference. Ydstie (1984) modified this control by introducing an extended horizon, applicable to nonminimum-phase systems, but not to open-loop unstable systems. The nearest control to GPC is predictive control by Peterka (1984) for an infinite prediction horizon, hence a different development. Clarke et al. (1987a,b) introduced generalized predictive control, which minimizes the following criterion

$$J(u,t) = \mathrm{E}\left\{ \sum_{j=N_1}^{N_2} [y(t+j) - r(t+j)]^2 + \sum_{j=1}^{N_u} \lambda(j)[\Delta u(t+j-1)]^2 \right\} \quad (15.4)$$

where N_1 and N_2 are the minimum and maximum cost horizons and N_u the control horizon (Fig. 15.1); these horizons are finite. $\lambda(j)$ is a weighting sequence for the input. In fact, only the first calculated input $u(t)$ is applied and the following inputs $u(t+1), \ldots$ are only calculated in open loop. The minimization is thus repeated at the following step. The calculation of the input necessitates us knowing the future set points, and the aim of the control is, owing to the output predictions, to bring the future outputs as close as possible to the set points. The most important parameter is the prediction horizon N_2. In general, the control horizon is chosen to be small.

15.3 Simple Generalized Predictive Control

15.3.1 Theoretical Presentation

In this section, the described method is the original method of Clarke et al. (1987a,b), restricted to the single-input single-output case.

The considered system model is the following

$$A(q^{-1})y(t) = B(q^{-1})u(t-1) + \frac{C(q^{-1})}{\Delta(q^{-1})}\xi(t) \quad (15.5)$$

based on the use of the backward shift operator q^{-1}. This is an auto-regressive integrated moving-average exogenous input (ARIMAX) model and $\xi(t)$ is white noise of zero mean. The polynomials are

$$\begin{array}{ll} A(q^{-1}) & = 1 + a_1\, q^{-1} + \ldots + a_{na}\, q^{-na} \\ B(q^{-1}) & = b_0 + b_1\, q^{-1} + \ldots + b_{nb}\, q^{-nb} \\ C(q^{-1}) & = 1 + c_1\, q^{-1} + \ldots + c_{nc}\, q^{-nc} \\ \Delta(q^{-1}) & = 1 - q^{-1} \end{array} \quad (15.6)$$

The natural time delay of the output with respect to the input is integrated in the model writing; if the process presents a zero time delay d, the first d elements of $B(q^{-1})$ are zero. Notice that the polynomials $A(q^{-1})$ and $C(q^{-1})$ are monic.

The model (15.5) can also be considered as

$$A(q^{-1})\,\Delta y(t) = B(q^{-1})\,\Delta u(t-1) + C(q^{-1})\,\xi(t) \quad (15.7)$$

making use of the variations in the input and the output.

We seek the inputs u minimizing the criterion (15.4). To solve this optimization problem, we need a j-step ahead predictor of the output (Favier and Dubois, 1990) depending on the past information and on the future input variations Δu which will be calculated, so that the criterion is optimal. Such a predictor is based on the solving of the Diophantine equation

$$C(q^{-1}) = E_j(q^{-1})\, A(q^{-1})\Delta(q^{-1}) + q^{-j}\, F_j(q^{-1}) \tag{15.8}$$

where the polynomials E_j and F_j only depend on $A(q^{-1})$, $C(q^{-1})$ and on degree j. Moreover, the degree of E_j is such that: $\deg(E_j) = j-1$. Equation (15.8) can be seen as the division of $C(q^{-1})$ by q^{-j} with quotient $F_j(q^{-1})$ and remaining $E_j(q^{-1})\, A(q^{-1})\Delta(q^{-1})$. This polynomial equation can be favourably solved in a recurrent manner (Acundeger and Favier, 1993), (Favier and Dubois, 1990). It is also possible to use a bank of predictors.

Using the model (15.7) and Eq. (15.8), we obtain

$$y(t+j) = \frac{B}{A}\, u(t+j-1) + E_j\, \xi(t+j) + \frac{F_j}{A\,\Delta}\xi(t) \tag{15.9}$$

which we still can transform by replacing the known terms $\xi(t)$ with respect to the model (15.5), and again using Eq. (15.8), hence

$$y(t+j) = \frac{F_j}{C}y(t) + \frac{E_j\, B}{C}\, \Delta u(t+j-1) + E_j\, \xi(t+j) \tag{15.10}$$

Noticing that the term $\xi(t+j)$ is independent of the past information at time t, the predictor results

$$\hat{y}(t+j) = \frac{F_j}{C}y(t) + \frac{E_j\, B}{C}\, \Delta u(t+j-1) \tag{15.11}$$

We then introduce a second Diophantine equation

$$E_j(q^{-1})\, B(q^{-1}) = G_j(q^{-1})\, C(q^{-1}) + q^{-j}\, \Gamma_j(q^{-1}) \tag{15.12}$$

to separate the past inputs from the future inputs (Bitmead et al., 1990). This is performed in the same way as in Eq. (15.8) with $\deg(G_j) = j-1$. The predictor expression (15.11) becomes

$$\hat{y}(t+j) = \frac{F_j}{C}y(t) + \frac{\Gamma_j}{C}\, \Delta u(t-1) + G_j(q^{-1})\, \Delta u(t+j-1) \tag{15.13}$$

which is written as

$$\hat{y}(t+j) = F_j y^f(t) + \Gamma_j\, u^f(t-1) + G_j(q^{-1})\, \Delta u(t+j-1) \tag{15.14}$$

by introducing the variations in the filtered input and the filtered output

$$\begin{array}{ll} u^f(t) & = 1/C(q^{-1})\, \Delta u(t) \\ y^f(t) & = 1/C(q^{-1})\, y(t) \end{array} \tag{15.15}$$

It is typical (Camacho and Bordons, 1998) to consider the control sequence formed by the sum of a free signal and a forced signal, which will induce the free and forced responses respectively. The free signal corresponds to the past inputs and is kept constant and equal to its last value at future time instants. The forced signal is taken to be equal to zero at past instants and equal to the control variations for future instants (Fig. 15.2).

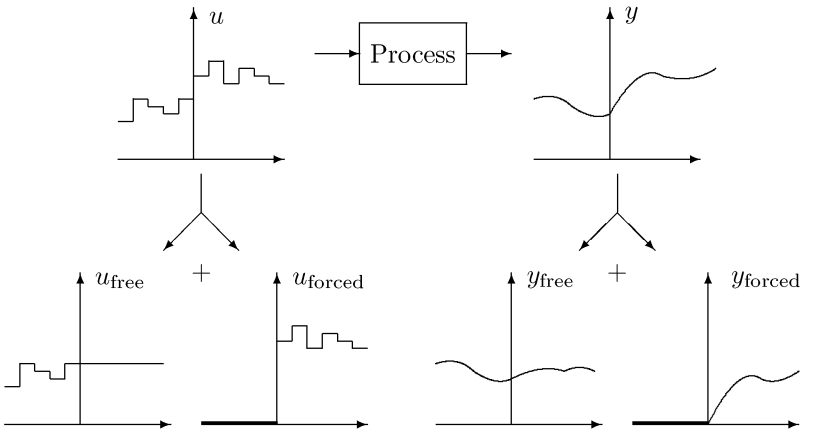


Figure 15.2: Decomposition of the control sequence into the sum of a free signal and a forced signal and decomposition of the resulting response into free and forced responses

From Eq. (15.14), results the prediction of the free response $\hat{y}(t+j|t)$ obtained by assuming that all the future input variations will be zero

$$\hat{y}(t+j|t) = F_j y^f(t) + \Gamma_j \, u^f(t-1) \tag{15.16}$$

so that the predictor is decomposed into two parts

$$\hat{y}(t+j) = \hat{y}(t+j|t) + G_j(q^{-1}) \, \Delta u(t+j-1) \tag{15.17}$$

The polynomials G_j and Γ_j can be calculated in a recurrent manner from Eq. (15.12). It can also be noticed that the coefficients g_i of polynomial $G_j(q^{-1})$ are the first j terms of the response to a step of the system model $B/(A\Delta)$, thus are the Markov parameters of this transfer function, as indicated by the following equation

$$\frac{B}{A\,\Delta} = G_j + q^{-j}\,\frac{\Gamma_j}{C} + q^{-j}\,\frac{B\,F_j}{A\,\Delta\,C} \tag{15.18}$$

obtained from both Diophantine Eqs. (15.8) and (15.12).

Defining the vector $\boldsymbol{f}$ of the predictions of the free response

$$\boldsymbol{f} = [\hat{y}(t+1|t), \ldots, \hat{y}(t+N_2|t)]^T \tag{15.19}$$

and the vector $\tilde{\boldsymbol{u}}$ of the future input variations

$$\tilde{\boldsymbol{u}} = [\Delta u(t), \ldots, \Delta u(t+N_u-1)]^T \tag{15.20}$$

the vector $\hat{\boldsymbol{y}}$ of predictions defined by

$$\hat{\boldsymbol{y}} = [\hat{y}(t+1), \ldots, \hat{y}(t+N_2)]^T \tag{15.21}$$

is equal to

$$\hat{\boldsymbol{y}} = \boldsymbol{G}\,\tilde{\boldsymbol{u}} + \boldsymbol{f} \tag{15.22}$$

where the matrix $\boldsymbol{G}$ of dimension $N_2 \times N_u$ contains as elements the first j coefficients g_i of polynomials G_j as

$$\boldsymbol{G} = \begin{bmatrix} g_0 & 0 & \ldots & 0 \\ g_1 & g_0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_{N_u-1} & g_{N_u-2} & \cdots & g_0 \\ \vdots & \vdots & \ldots & \vdots \\ g_{N_2-1} & g_{N_2-2} & \cdots & g_{N_2-N_u} \end{bmatrix} \tag{15.23}$$

If the process time delay is larger than the natural delay: $d > 1$, the first $d-1$ rows of $\boldsymbol{G}$ are zero. It is possible to choose $N_1 = d$ to avoid it. Nevertheless, Clarke et al. (1987a) notice that GPC provides a stable solution even when the first $d-1$ rows of $\boldsymbol{G}$ are zero.

The quadratic criterion to be minimized becomes

$$\begin{aligned} J(u,t) &= \mathrm{E}\left\{(\boldsymbol{y}-\boldsymbol{r})^T(\boldsymbol{y}-\boldsymbol{r}) + \lambda\,\tilde{\boldsymbol{u}}^T\tilde{\boldsymbol{u}}\right\} \\ &= (\boldsymbol{G}\,\tilde{\boldsymbol{u}} + \boldsymbol{f} - \boldsymbol{r})^T(\boldsymbol{G}\,\tilde{\boldsymbol{u}} + \boldsymbol{f} - \boldsymbol{r}) + \lambda\,\tilde{\boldsymbol{u}}^T\tilde{\boldsymbol{u}} \end{aligned} \tag{15.24}$$

assuming the series $\lambda(j)$ is constant and setting the reference vector $\boldsymbol{r}$ equal to

$$\boldsymbol{r} = [r(t+1), \ldots, r(t+N_2)]^T \tag{15.25}$$

Assuming that no constraints on the input exist, we obtain the vector of the future input variations, solution of this problem of minimization of the cost criterion

$$\tilde{\boldsymbol{u}} = [\boldsymbol{G}^T\boldsymbol{G} + \lambda\,\boldsymbol{I}]^{-1}\,\boldsymbol{G}^T(\boldsymbol{r} - \boldsymbol{f}) \tag{15.26}$$

In fact, only the first input is really applied; thus, we obtain

$$u(t) = u(t-1) + \bar{\boldsymbol{g}}^T(\boldsymbol{r} - \boldsymbol{f}) \tag{15.27}$$

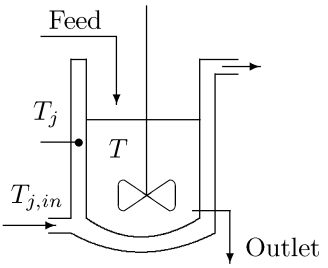
where $\bar{\boldsymbol{g}}$ is the first row of matrix $[\boldsymbol{G}^T\boldsymbol{G} + \lambda\,\boldsymbol{I}]^{-1}\,\boldsymbol{G}^T$.

According to Clarke et al. (1987a), the control horizon N_u is a very important parameter. Nevertheless, it is often sufficient to take it to be equal to 1, except for complex systems where N_u will have to be chosen to be equal to the number of unstable or badly damped poles. To bring more damping, λ must be increased. GPC can be used for a nonminimum-phase system, even in the case where $\lambda = 0$.

The relation between GPC and pole-placement control as well as linear quadratic control was studied by Clarke et al. (1987a,b) and particularly by Bitmead et al. (1990).

15.3.2 Numerical Example : Generalized Predictive Control of a Chemical Reactor

Example 15.1: Generalized Predictive Control of a Chemical Reactor



The studied case concerns the chemical reactor described in Sect. 19.2. The identification was realized by recursive least squares, taking Δu and Δy as components of the observation vector, in order to model the process by an ARIMAX model of Eq. (15.7).

The polynomials A and B, whose coefficients have been simplified to facilitate the numerical demonstration, are equal to

$$A(q^{-1}) = 1 - 0.97\,q^{-1} \quad ; \quad B(q^{-1}) = 1.2 + 0.58\,q^{-1} \tag{15.28}$$

while C is equal to 1.

In the case where the control horizon N_u and the prediction horizon are both equal to 3, we obtain, as a solution of the first Diophantine Eq. (15.8), the following polynomials E_j and F_j

$$\begin{array}{lll} j=1 & E_1 = 1 & F_1 = 1.97 - 0.97\,q^{-1} \\ j=2 & E_2 = 1 + 1.97\,q^{-1} & F_2 = 2.9109 - 1.9109\,q^{-1} \\ j=3 & E_3 = 1 + 1.97\,q^{-1} + 2.9109\,q^{-2} & F_3 = 3.8236 - 2.8236\,q^{-1} \end{array} \tag{15.29}$$

hence the predictors according to Eq. (15.11)

$$\begin{aligned} \hat{y}(t+1) =& 1.97y(t) - 0.97y(t-1) + 1.2\Delta u(t) + 0.58\Delta u(t-1) \\ \hat{y}(t+2) =& 2.9109y(t) - 1.9109y(t-1) \\ & +1.2\Delta u(t+1) + 2.944\Delta u(t) + 1.4426\Delta u(t-1) \\ \hat{y}(t+3) =& 3.8236y(t) - 2.8236y(t-1) \\ & +1.2\Delta u(t+2) + 2.944\Delta u(t+1) + 4.6357\Delta u(t) + 1.6883\Delta u(t-1) \end{aligned} \tag{15.30}$$

For the second Diophantine Eq. (15.12), the polynomials G_j and Γ_j are

$$\begin{array}{lll} j=1 & G_1 = 1.2 & \Gamma_1 = 0.58 \\ j=2 & G_2 = 1.2 + 2.944\,q^{-1} & \Gamma_2 = 1.1426 \\ j=3 & G_3 = 1.2 + 2.944\,q^{-1} + 4.6357\,q^{-2} & \Gamma_3 = 1.6883 \end{array} \tag{15.31}$$

The predictors of the free responses are calculated from Eq. (15.16)

$$\begin{aligned} \hat{y}(t+1|t) &= 0.58\Delta u(t-1) + 1.97y(t) - 0.97y(t-1) \\ \hat{y}(t+2|t) &= 1.4426\Delta u(t-1) + 2.9109y(t) - 1.9109y(t-1) \\ \hat{y}(t+3|t) &= 1.6883\Delta u(t-1) + 3.8236y(t) - 2.8236y(t-1) \end{aligned} \tag{15.32}$$

The matrix G is equal to

$$G = \begin{bmatrix} 1.2 & 0 & 0 \\ 2.944 & 1.2 & 0 \\ 4.6357 & 2.944 & 1.2 \end{bmatrix} \tag{15.33}$$

and the gain matrix of the control law (for $\lambda = 0.1$)

$$[G^T G + \lambda I]^{-1} G^T = \begin{bmatrix} 0.5181 & 0.1823 & -0.0435 \\ -1.0888 & 0.1876 & 0.1823 \\ 0.6262 & -1.0888 & 0.5181 \end{bmatrix} \tag{15.34}$$

so that the control law in an explicit form is

$$\begin{aligned} \Delta u(t) &= 0.5181[r(t+1) - \hat{y}(t+1|t)] + 0.1823[r(t+2) - \hat{y}(t+2|t)] \\ &\quad -0.0435[r(t+3) - \hat{y}(t+3|t)] \\ &= -0.490\Delta u(t-1) - 1.385y(t) + 0.7281y(t-1) \\ &\quad +0.5181r(t+1) + 0.1823r(t+2) - 0.0435r(t+3) \end{aligned} \tag{15.35}$$

The polynomials R, S and T corresponding to the pole-placement equivalent of this control (Fig. 15.3) are, approximately

$$\begin{aligned} &R = 1.385 - 0.728q^{-1} \quad ; \quad S = 1 + 0.4354q^{-1} \\ &T = 0.5181q^{-2} + 0.1823q^{-1} - 0.0435 \end{aligned} \tag{15.36}$$

The simulation results are given and commented on in Sect. 15.6.

15.3.3 GPC seen as a Pole-Placement

As only the first row of matrix $[G^T G + \lambda I]^{-1} G^T$ is used for the effective control $u(t)$, the elements of this row are particularized and will be denoted by $\bar{g}_i$ in this section ($i = 1, \ldots, N_2$). From Eq. (15.26), we deduce the control variation

$$\begin{aligned} \Delta u(t) &= \sum_{i=1}^{N_2} \bar{g}_i \, \{r(t+i) - \hat{y}(t+i|t)\} \\ &= \sum_{i=1}^{N_2} \bar{g}_i \left\{ -\frac{\Gamma_i(q^{-1})}{C(q^{-1})} \Delta u(t-1) + r(t+i) - \frac{F_i(q^{-1})}{C(q^{-1})} y(t) \right\} \end{aligned} \tag{15.37}$$

which we can transform into the classical RST form of pole-placement (Fig. 15.3)

$$\begin{aligned} \left\{ C(q^{-1}) + \sum_{i=1}^{N_2} \bar{g}_i q^{-1} \Gamma_i(q^{-1}) \right\} \Delta u(t) = & \left\{ C(q^{-1}) \sum_{i=1}^{N_2} \bar{g}_i q^{-N_2+i} \right\} r(t+N_2) \\ & - \left\{ \sum_{i=1}^{N_2} \bar{g}_i F_i \right\} y(t) \iff \\ S\Delta u(t) = & \, C(q^{-1}) T r(t+N_2) - R y(t) \end{aligned} \tag{15.38}$$

by setting the polynomials R, S, T of the pole-placement

$$\begin{aligned} R &= \sum_{i=1}^{N_2} \bar{g}_i F_i \\ S &= C(q^{-1}) + \sum_{i=1}^{N_2} \bar{g}_i q^{-1} \Gamma_i(q^{-1}) \\ T &= \sum_{i=1}^{N_2} \bar{g}_i q^{-N_2+i} \end{aligned} \tag{15.39}$$

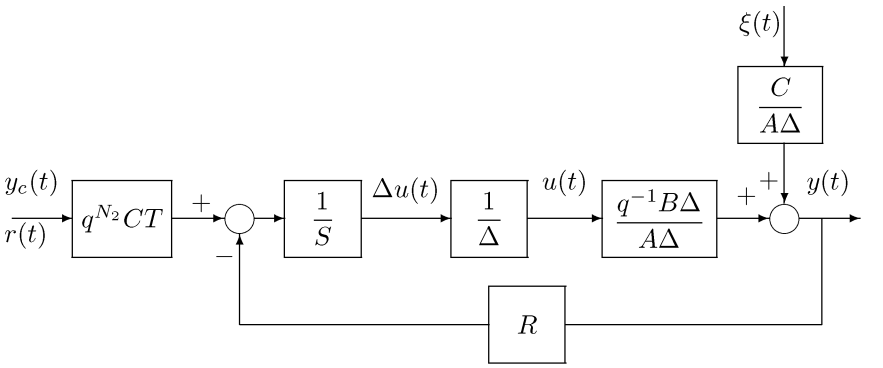


Figure 15.3: Generalized predictive control considered as pole-placement

GPC can thus be presented as a particular pole-placement minimizing a criterion. Using the process model (15.5), we have

$$(A\Delta S + q^{-1}BR)y(t) = BCTr(t + N_2 - 1) + CS\xi(t) \tag{15.40}$$

as on the other hand, by expressing R and S, and using Eq. (15.18), we have

$$A\Delta S + q^{-1}\,BR = C\left\{A\Delta + \sum_{i=1}^{N_2} \bar{g}_i q^{i-1}\,(B - A\Delta G_i)\right\} \tag{15.41}$$

and we obtain

$$\begin{aligned} \underbrace{\left\{A\Delta + \sum_{i=1}^{N_2} \bar{g}_i q^{i-1}\,(B - A\Delta G_i)\right\}}_{A_c} y(t) &= BTr(t + N_2 - 1) + S\xi(t) \\ A_c y(t) &= BTr(t + N_2 - 1) + S\xi(t) \end{aligned} \tag{15.42}$$

which shows that the closed-loop output does not depend on polynomial C. The stability condition is that the polynomial A_c has its roots inside the unit circle. The influence of parameters N_1, N_2, N_u, λ is complex. It is easy to verify that the steady-state gain of the closed-loop transfer function is equal to 1.

15.4 Generalized Predictive Control with Multiple Reference Model

15.4.1 Theoretical Presentation

This variant of GPC was introduced by Irving et al. (1986). It introduces a different specification for the regulation dynamics and the tracking dynamics, and thus allows us to improve the robustness.

On the other hand, the reference model is given by the equation

$$r(t) = \frac{B(q^{-1})T}{A_m(q^{-1})} y_c(t) \tag{15.43}$$

where y_c is the set point, i.e the input of the reference model, $A_m(q^{-1})$ the stable monic polynomial fixing the tracking dynamics and T a scalar ensuring a unit gain.

On the other hand, define $u_r(t)$ as the control which would give an output $y(t)$ equal to the reference $r(t)$ in the absence of noise. From the system transfer function $q^{-1}B(q^{-1})/A(q^{-1})$, we deduce

$$u_r(t) = \frac{A(q^{-1})T}{A_m(q^{-1})} y_c(t+1) \tag{15.44}$$

Then, we introduce the respective deviations of the input and the output with respect to their reference

$$\begin{aligned} e^u(t) &= \Delta(u(t) - u_r(t)) \\ e^y(t) &= y(t) - r(t) \end{aligned} \tag{15.45}$$

These deviations are filtered by a stable monic polynomial $A_f(q^{-1})$, which modifies the regulation dynamics

$$\begin{aligned} e^{uf}(t) &= A_f(q^{-1})e^u(t) \\ e^{yf}(t) &= A_f(q^{-1})e^y(t) \end{aligned} \tag{15.46}$$

The optimization criterion (15.4) is replaced by a criterion depending on the filtered deviations

$$J(u,t) = \mathrm{E}\left\{ \sum_{j=N_1}^{N_2} [e^{yf}(t+j)]^2 + \sum_{j=1}^{N_u} \lambda(j)[e^{uf}(t+j-1)]^2 \right\} \tag{15.47}$$

with: $e^{uf}(t+j) = 0$ $(i = N_u, N_2)$. The method is completely parallel to simple GPC where the process model is replaced by the following performance model

$$A\Delta e^{yf} = Be^{uf}(t-1) + A_f C\xi(t) \tag{15.48}$$

The developed equations will be very close. The two Diophantine equations to solve are

$$\begin{aligned} A_f(q^{-1})\, C(q^{-1}) &= E_j(q^{-1})\, A(q^{-1})\, \Delta(q^{-1}) + q^{-j}\, F_j(q^{-1}) \\ E_j(q^{-1})\, B(q^{-1}) &= G_j(q^{-1})\, A_f(q^{-1})\, C(q^{-1}) + q^{-j}\, \Gamma_j(q^{-1}) \end{aligned} \tag{15.49}$$

The predictions are expressed with respect to deviations

$$\hat{e}^{yf}(t+j) = \frac{F_j}{A_f C} e^{yf}(t) + \frac{\Gamma_j}{A_f C}\, e^{uf}(t-1) + G_j(q^{-1})\, e^{uf}(t+j-1) \quad (15.50)$$

giving the predictions of the deviations of the free responses

$$\hat{e}^{yf}(t+j) = \frac{F_j}{A_f C} e^{yf}(t) + \frac{\Gamma_j}{A_f C}\, e^{uf}(t-1) \quad (15.51)$$

The following vectors are then defined

$$\begin{array}{l} \boldsymbol{f} = [\hat{e}^{yf}(t+1|t), \ldots, \hat{e}^{yf}(t+N_2|t)]^T \\ \boldsymbol{e^u} = [e^{uf}(t), \ldots, e^{uf}(t+N_u-1)]^T \\ \boldsymbol{\hat{e}^y} = [\hat{e}^{yf}(t+1), \ldots, \hat{e}^{yf}(t+N_2)]^T \end{array} \quad (15.52)$$

These vectors are related by the relation

$$\boldsymbol{\hat{e}^y} = \boldsymbol{G}\, \boldsymbol{e^u} + \boldsymbol{f} \quad (15.53)$$

and the quadratic criterion to minimize is

$$J = \boldsymbol{\hat{e}^{y}}^T \boldsymbol{\hat{e}^y} + \lambda \boldsymbol{e^u}^T \boldsymbol{e^u} \quad (15.54)$$

hence the solution vector

$$\boldsymbol{e^u} = -[\boldsymbol{G}^T \boldsymbol{G} + \lambda\, \boldsymbol{I}]^{-1}\, \boldsymbol{G}^T\, \boldsymbol{f} \quad (15.55)$$

whose first component only is applied for the real control. With the matrix $\boldsymbol{G}$ being the same as for GPC, the elements $\bar{g}_i$ are identical and we obtain

$$e^{uf}(t) = -\sum_{j=1}^{N_2} \bar{g}_j\, \hat{e}^{yf}(t+j|t) \quad (15.56)$$

which can be expressed as

$$\begin{array}{ll} (A_f C + \sum_{j=1}^{N_2} \bar{g}_j\, q^{-1}\, \Gamma_j)\Delta(u(t) - u^r(t)) & = -\sum_{j=1}^{N_2} \bar{g}_j\, F_j(y(t) - r(t)) \Longleftrightarrow \\ S\Delta(u(t) - u^r(t)) & = -R(y(t) - r(t)) \Longleftrightarrow \\ S e^{uf}(t) & = -R e^{yf}(t) \end{array} \quad (15.57)$$

Again, GPC with a reference model can be considered as a particular pole-placement minimizing a criterion (Fig. 15.4), by simply setting

$$\begin{array}{ll} S & = A_f C + \sum_{j=1}^{N_2} \bar{g}_j\, q^{-1}\, \Gamma_j \\ R & = \sum_{j=1}^{N_2} \bar{g}_j\, F_j \end{array} \quad (15.58)$$

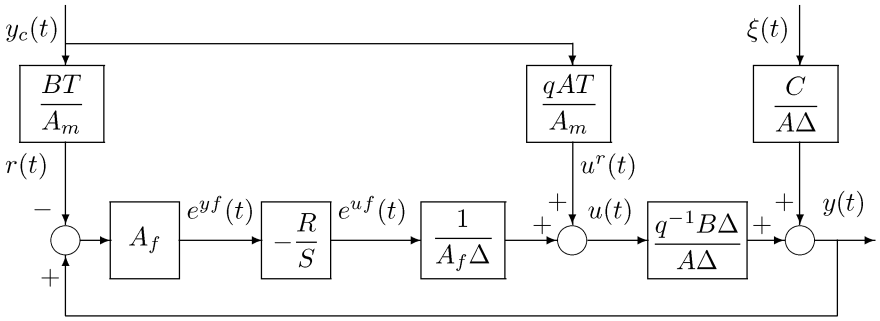


Figure 15.4: GPC with performance model seen as a pole-placement

The closed-loop equation is very similar to Eq. (15.40) and is equal to

$$(A\Delta S + q^{-1}BR)y(t) = (A\Delta S + q^{-1}BR)r(t + N_2 - 1) + CS\xi(t) \tag{15.59}$$

thus we can draw

$$y(t) = \frac{BT}{A_m}y_c(t) + \frac{CS}{A\Delta S + q^{-1}BR}\xi(t) \tag{15.60}$$

which displays the stability condition that the polynomial $A\Delta S + q^{-1}BR$ has its roots inside the unit circle. In fact, after development of R and S, and use of both Diophantine equations, this polynomial can be expressed as

$$\begin{aligned} A\Delta S + q^{-1}BR &= A_f\, C\, [A\Delta + \sum_{j=1}^{N_2} \bar{g}_j q^{j-1}(B - A\Delta G_j)] \\ &= A_f\, C\, A_c \end{aligned} \tag{15.61}$$

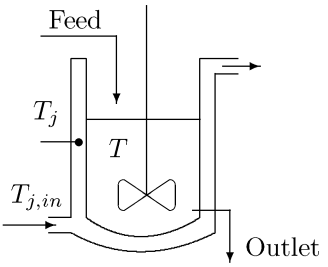
The output is thus

$$y(t) = \frac{BT}{A_m}y_c(t) + \frac{S}{A_f\, A_c}\xi(t) \tag{15.62}$$

which emphasizes the role played by the polynomial A_f acting on the regulation dynamics, and by A_m acting on the tracking dynamics. Note that polynomial A_c depends, in fact, only on matrix $\boldsymbol{G}$ and parameter λ. It is advisable to choose the parameters so that the dynamics of polynomial A_c is negligible (fast) compared to those of A_f and A_m.

15.4.2 Numerical Example: Generalized Predictive Control with Performance Model of a Chemical Reactor

Example 15.2: **Generalized Predictive Control with Performance Model of a Chemical Reactor**



Again, consider the previously studied case, concerning the chemical reactor, in order to evaluate the influence of the introduction of the performance model on GPC.

The reactor model remains the same with

$$A(q^{-1}) = 1 - 0.97\,q^{-1} \quad ; \quad B(q^{-1}) = 1.2 + 0.58\,q^{-1} \tag{15.63}$$

while C is equal to 1.

We choose the following polynomial A_m for the reference trajectory

$$A_m(q^{-1}) = 1 - 0.82\,q^{-1} \tag{15.64}$$

so as to provide a smooth trajectory, whereas the constant T ensures the steady-state gain of $B(q^{-1})T/A_m(q^{-1})$ equal to 1, hence $T = A_m(1)/B(1)$.

The polynomial influencing the regulation is taken to be equal to

$$A_f(q^{-1}) = 1 - 0.9\,q^{-1} \tag{15.65}$$

The control horizon N_u and the prediction horizon are both kept equal to 3. We obtain as a solution of the first Diophantine Eq. (15.49) the following polynomials E_j and F_j

$$\begin{array}{lll} j=1 & E_1 = 1 & F_1 = 1.07 - 0.97\,q^{-1} \\ j=2 & E_2 = 1 + 1.07\,q^{-1} & F_2 = 1.1379 - 1.0379\,q^{-1} \\ j=3 & E_3 = 1 + 1.07\,q^{-1} + 1.1379\,q^{-2} & F_3 = 1.2038 - 1.1038\,q^{-1} \end{array} \tag{15.66}$$

For the second Diophantine Eq. (15.49), the polynomials G_j (unchanged compared to the case without performance model) and Γ_j are the following

$$\begin{array}{lll} j=1 & G_1 = 1.2 & \Gamma_1 = 1.660 \\ j=2 & G_2 = 1.2 + 2.944\,q^{-1} & \Gamma_2 = 3.2702 \\ j=3 & G_3 = 1.2 + 2.944\,q^{-1} + 4.6357\,q^{-2} & \Gamma_3 = 4.8321 \end{array} \tag{15.67}$$

The matrix G is unchanged, thus also the matrix $[G^T G + \lambda I]^{-1} G^T$. The polynomials R and S of the equivalent pole-placement of this control (Fig. 15.4) are approximately

$$R = 0.709 - 0.644q^{-1} \quad ; \quad S = 1 + 0.346q^{-1} \quad ; \quad T = \frac{A_m(1)}{B(1)} = 0.101 \tag{15.68}$$

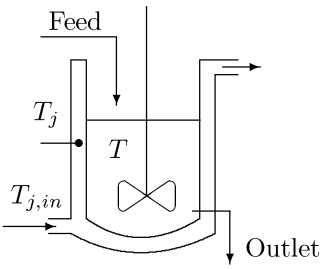
The simulation results are given and commented on in Sect. 15.6.

15.5 Partial State Reference Model Control

(M'Saad et al., 1990) extended GPC with reference model under the name of partial state reference model adaptive control (Corriou, 1996). Many applications have been realized (Bendotti and Msaad, 1993), (M'saad et al., 1993), including multivariable control (M'saad and Sanchez, 1992), (M'saad and Sanchez, 1994). Using a reference model, this technique allows us to separately specify the regulation and tracking dynamics. In the case of a criterion with finite horizons, the approach is similar to that of GPC, whereas in the case of infinite prediction and control horizons, the proposed controller is based on infinite horizon linear quadratic control, which is stable provided that the process model is observable and stabilizable, while the stability of GPC is linked to the synthesis parameters. Moreover, with infinite horizons, it is sufficient to specify only one parameter against four parameters for GPC.

15.6 Generalized Predictive Control of a Chemical Reactor

Example 15.3: Generalized Predictive Control of a Chemical Reactor



The temperature of the chemical reactor contents is controlled by manipulating the position of a three-way valve, connected to two heat exchangers that are hot and cold respectively, and allowing action on the temperature of the heat-conducting fluid entering the jacket. The reactor model corresponds to the reactor described in Sect. 19.2 and considered in both the previous numerical examples.

In this simulated reactor, GPC is tested for different values of the prediction horizon N_2 and the control horizon N_1. The parameter λ is maintained equal to 0.1.

Figures 15.5 to 15.7, have been obtained for the same prediction horizon N_2 and control horizons N_u increasing from 1 to 3. The response quality is slightly improved when N_u increases, although it was already very correct for $N_u = 1$. The peak of control u increases at the instants corresponding to the set point changes. Figure 15.8 has been obtained for minima horizons $N_2 = 1, N_u = 1$. The control input u presents sharper peaks than when $N_2 = 3$.

The disturbance rejection is tested by imposing a step disturbance of amplitude 0.2 between $t = 20$ s and $t = 30$ s (Fig. 15.9), with the horizons $N_2 = 3$ and $N_u = 3$. It is clear that the disturbance is efficiently rejected.

Gaussian white noise of standard deviation 0.2 was added to the output (Fig. 15.10). In spite of this noise, the output follows, in average, the set point

without problem.

GPC with different reference models for regulation and tracking has been tested. The influence of the introduction of these models appears neatly by comparing the new Fig. 15.11 to a similar case, but without models (Fig. 15.7). The input varies more smoothly, while the output follows very regularly its set point trajectory, not represented in the figure (because it is superposed). Again, in the case with reference models, the influence of white noise on the output is less sensitive (Fig. 15.12).

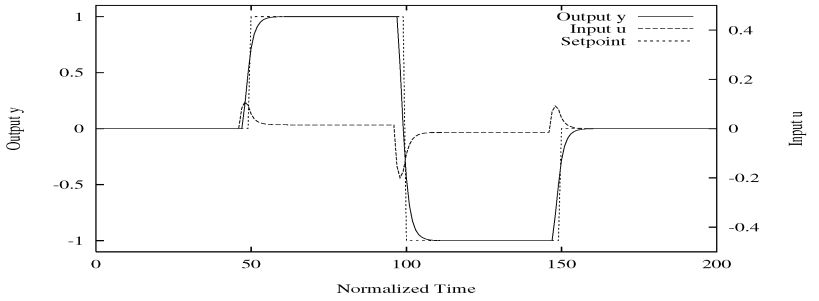


Figure 15.5: GPC of the identified linear model of the chemical reactor ($N_1 = 1, N_2 = 3, N_u = 1$). Variations in the controlled output and the control input

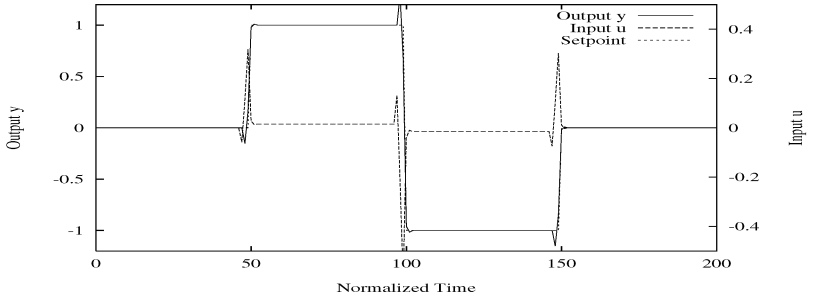


Figure 15.6: GPC of the identified linear model of the chemical reactor ($N_1 = 1, N_2 = 3, N_u = 2$). Variations in the controlled output and the control input

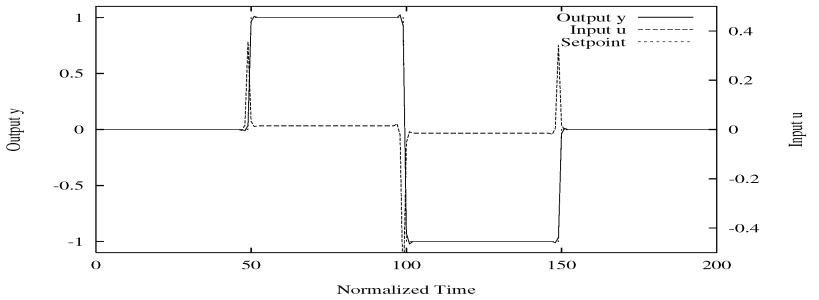


Figure 15.7: Generalized predictive control of the identified linear model of the chemical reactor ($N_1 = 1, N_2 = 3, N_u = 3$). Variations of the controlled output and of the control input

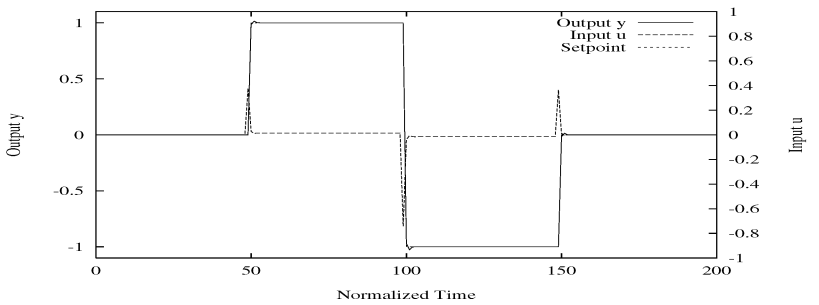


Figure 15.8: GPC of the identified linear model of the chemical reactor ($N_1 = 1, N_2 = 1, N_u = 1$). Variations in the controlled output and the control input

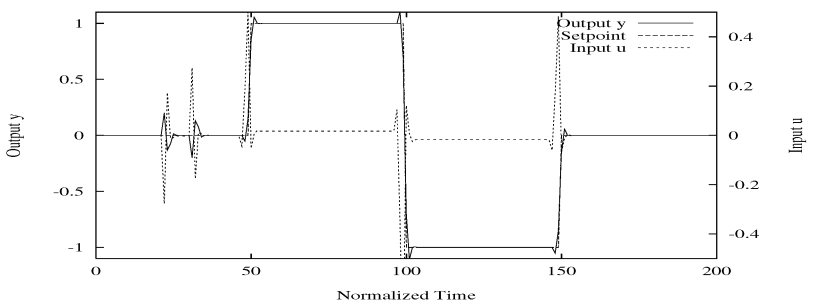


Figure 15.9: GPC of the identified linear model of the chemical reactor ($N_1 = 1, N_2 = 3, N_u = 3$) with step disturbance of amplitude 0.2 between $t = 20$ and $t = 30$. Variations in the controlled output and the control input

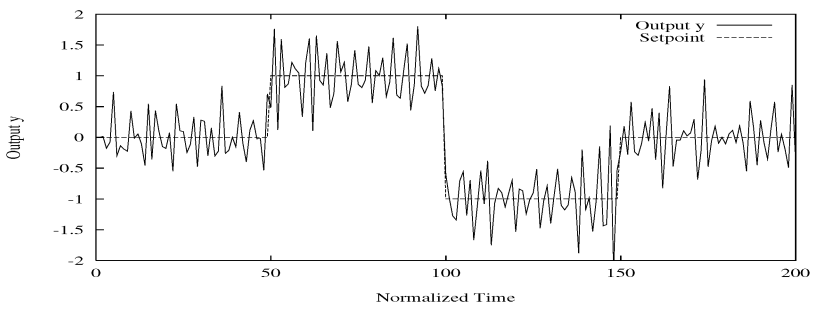


Figure 15.10: GPC of the identified linear model of the chemical reactor ($N_1 = 1, N_2 = 3, N_u = 3$) with Gaussian white noise on the output of standard deviation 0.2. Variations in the controlled output

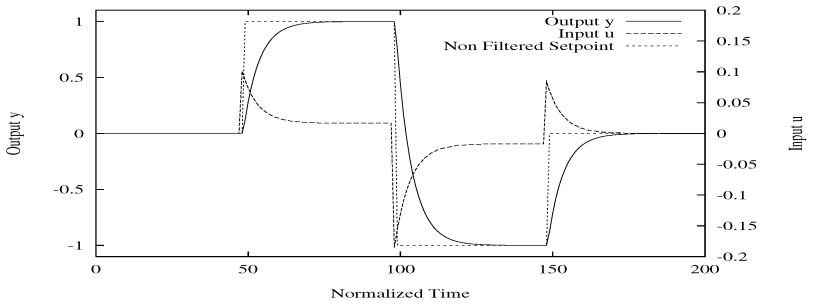


Figure 15.11: GPC of the identified linear model of the chemical reactor ($N_1 = 1, N_2 = 3, N_u = 3$) with reference model. Variations in the controlled output and the control input

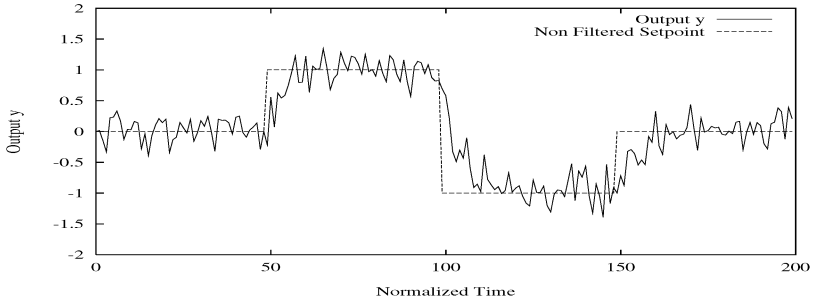


Figure 15.12: GPC of the identified linear model of the chemical reactor ($N_1 = 1, N_2 = 3, N_u = 3$) with reference model and with Gaussian white noise on the output of standard deviation 0.2. Variations in the controlled output

References

E. Acundeger and G. Favier. New computation formulae for k-step ahead predictors. In *European Control Conference- Groningen*, pages 951–957, 1993.

K.J. Aström. *Introduction to Stochastic Control Theory.* Academic Press, New York, 1970.

K.J. Aström and B. Wittenmark. On self tuning regulators. *Automatica*, pages 185–199, 1973.

P. Bendotti and M. Msaad. A skid-to-turn missile autopilot design: the generalized predictive adaptive control approach. *International Journal of Adaptive Control and Signal Processing*, 7:13–31, 1993.

R. R. Bitmead, M. Gevers, and V. Wertz. *Adaptive Optimal Control, The Thinking Man's GPC.* Prentice Hall, New York, 1990.

E.F. Camacho and C. Bordons. *Model Predictive Control.* Springer-Verlag, Berlin, 1998.

D.W. Clarke. Application of generalized predictive control to industrial processes. *IEEE Control Magazine*, 8:49–55, 1988.

D.W. Clarke and P.J. Gawthrop. Self-tuning controller. *Proc. IEE*, 122:929–934, 1975.

D.W. Clarke and P.J. Gawthrop. Self-tuning control. *Proc. IEE*, 123:633–640, 1979.

D.W. Clarke, C. Mohtadi, and P.S. Tuffs. Generalized predictive control - Part I. The basic algorithm. *Automatica*, 23(2):137–148, 1987a.

D.W. Clarke, C. Mohtadi, and P.S. Tuffs. Generalized predictive control - Part II. Extensions and interpretations. *Automatica*, 23(2):149–160, 1987b.

J.P. Corriou. *Commande des Procédés.* Lavoisier, Tec. & Doc., Paris, 1996.

G. Defaye, N. Regnier, J. Chabanon, L. Caralp, and C. Vidal. Adaptive-predictive temperature control of semi-batch reactors. *Chem. Eng. Sci.*, 48 (19):3373–3382, 1993.

G. Favier and D. Dubois. A review of k-step-ahead predictors. *Automatica*, 26 (1):75–84, 1990.

E. Irving, C.M. Falinower, and C. Fonte. Adaptive generalized predictive control with multiple reference models. In *2nd IFAC Workshop on adaptive systems in control and signal processing, Lund, Sweden*, 1986.

M.V. LeLann, K. Najim, and G. Casamatta. Generalized predictive control of a pulsed liquid-liquid extraction column. *Chem. Eng. Comm.*, 48:237–253, 1986.

M. M'saad and G. Sanchez. Partial state reference model adaptive control of multivariable systems. *Automatica*, 28(6):1189–1197, 1992.

M. M'saad and G. Sanchez. Multivariable generalized predictive adaptive control with a suitable tracking capability. *J. Proc. Cont.*, 4(1):45–52, 1994.

M. M'Saad, I.D. Landau, and M. Samaan. Further evaluation of partial state model reference adaptive design. *International Journal of Adaptive Control and Signal Processing*, 4:133–148, 1990.

M. M'saad, L. Dugard, and S. Hammad. A suitable generalized predictive adaptive controller case study: Control of a flexible arm. *Automatica*, 3: 589–608, 1993.

V. Peterka. Predictor-based self-tuning control. *Automatica*, 20:39–50, 1984.

A. Rafilamanana, M. Cabassud, M.V. LeLann, and G. Casamatta. Adaptive control of a multipurpose and flexible semi-batch pilot plant reactor. *Comp. Chem. Engng.*, 16(9):837–848, 1992.

J.B. Rawlings, E.S. Meadows, and K.R. Muske. Nonlinear model predictive control: A tutorial and survey. In *Advanced Control of Chemical Processes*, pages 203–224, Kyoto (Japan), 1994. IFAC.

B.E. Ydstie. Extended horizon adaptive control. pages 911–915, Oxford, 1984. IFAC 9th World Congress Budapest Hungary, Pergamon.

Chapter 16

Model Predictive Control

Model predictive control (MPC) is widely used in the industry, and many references to industrial experience will serve to present the main characteristics of this important control approach. Furthermore, the method of dynamic matrix control, about which much literature is available, will be discussed with some detail. Finally, some general aspects of nonlinear model predictive control will be outlined.

16.1 A General View of Model Predictive Control

A general objective of control schemes is to maintain the controlled variables close to their set points while respecting the process operating constraints. MPC has been designed to meet those purposes. It was first introduced by Richalet et al. (1978) (Adersa) as model algorithmic control (MAC) by IDCOM (IDentification-COMmand), where the accent was placed on the key role of digital computation and modelling; several industrial applications were reported to emphasize the interest of the proposed method. Very soon after, dynamic matrix control (DMC) was published (Cutler and Ramaker, 1979) and implemented at Shell as a multivariable computer control algorithm.

Many review papers have been devoted to the evolution of MPC (Garcia et al., 1989; Keyser et al., 1988; Lee, 1996; Mayne, 1996; Morari and Lee, 1991, 1999; Muske and Rawlings, 1993; Rawlings et al., 1994; Ricker, 1991), including books (Bitmead et al., 1990; Camacho and Bordons, 1995; Clarke, 1994; Maciejowski, 2002; Sanchez and Rodellar, 1996; Soeterboek, 1992) and several papers that relate industrial developments and applications of MPC (Froisy, 1994; Qin and Badgwell, 1996). Soeterboek (1992), in particular, through a unified approach compares different types of MPC schemes such as DMC (Cutler and Ramaker, 1979), PCA, AC, GPC (Clarke et al., 1987a,b), EPSAC (extended prediction self-adaptive control) (Keyser and Cauwenberghe, 1985) and extended horizon adaptive control (EHAC) (Ydstie, 1984).

IMC (Garcia and Morari, 1982) studied in different forms in chapters 4, 8 and 13, can also be considered as a variant of MPC.

GPC, which is a single-input single-output form of MPC was studied in Chap. 15.

Infinite horizon control as linear quadratic (LQ) and linear quadratic gaussian (LQG) control has been studied in Chap. 14.

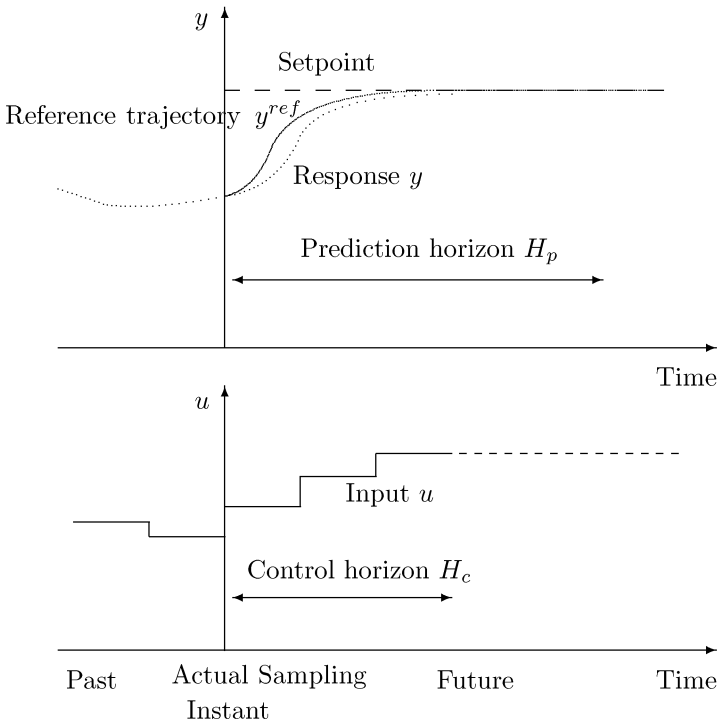


Figure 16.1: Principle of model predictive control with prediction and control horizons

MPC can be defined as a class of control algorithms which compute over a future time horizon (Fig. 16.1) a sequence of manipulated variables profiles by using a linear or nonlinear model of the plant in order to optimize a generally quadratic criterion subject to linear or nonlinear constraints.

Richalet et al. (1978) very clearly explained the main interest in an optimization control algorithm. They distinguished four hierarchical levels (from lower to higher):

- Level 0: control of ancillary systems (e.g. valves) where PID controllers are efficient.
- Level 1: dynamic control of the plant as a multivariable process perturbed by state and structural non-measured perturbations.
- Level 2: optimization of the set-points with minimization of cost functions,

ensuring quality and quantity of production.
- Level 3: time and space scheduling of production (planning-operation research).

As the economic benefits of levels 0 and 1 are low compared to those of level 2, effort should be concentrated on level 2 which can be defined as dynamic optimization or open-loop optimal control and concerns the calculation of the optimal trajectories to be followed. Of course, robustness aspects should be taken into account (Genceli and Nikolaou, 1993; Vuthandam et al., 1995), but the main characteristics are well defined.

Qin and Badgwell (1996) describes the general objectives of MPC in decreasing order of importance:

1. Prevent violation of input and output constraints.
2. Drive the manipulated variables towards their steady-state optimal values (dynamic input optimization).
3. Drive the controlled variables towards their steady-state optimal values taking into account the remaining degrees of freedom (dynamic output optimization).
4. Avoid excessive variation of manipulated variables.
5. When signals and actuators fail, control as much of the plant as possible.

The commercial codes for MPC adapt themselves in different ways to these rules. A flow chart for MPC calculation is given by Qin and Badgwell (1996) which follows the stages:

- Read the manipulated variables, disturbances and controlled variables of the process at time k.
- Update the outputs by means of feedback. A bias between the current measured and the current predicted output is defined:

$$\boldsymbol{b}_k = \boldsymbol{y}_k^m - \hat{\boldsymbol{y}}_k \tag{16.1}$$

  and this bias is added to the output model (16.5) for the following predictions

$$\hat{\boldsymbol{y}}_{k+j} = \boldsymbol{g}(\boldsymbol{x}_{k+j}) + \boldsymbol{b}_k \tag{16.2}$$

  This allows us to eliminate steady-state offset.
- Determine the controlled subprocess or which inputs must be manipulated and which outputs must be controlled by examining the degrees of freedom (Froisy, 1994): some outputs cannot be controlled if there are not enough manipulated variables; the system is square if there are the same number of manipulated and controlled variables; when there are more manipulated variables than controlled variables, it results in extra degrees of freedom, thus the optimization yields several solutions and another subsequent optimization can be performed (example of two-stage optimization in IDCOM-M at Setpoint and HIECON at Adersa).
- Remove ill-conditioning measured by the condition number of $\boldsymbol{G}^T\boldsymbol{G}$ where $\boldsymbol{G}$ is the process gain matrix. Different approaches are used in commercial codes, including the use of a threshold for singular values (RMPCT of Honeywell) and neglecting smaller singular values, or controlled variables controllability ranks defined by the user (IDCOM).

- Perform a local steady-state optimization by various means according to the codes: use a steady-state nonlinear model of the plant, use a linearized model, use the steady-state optimization problem with constraints.
- Perform the dynamic optimization, which consists of solving the following problem[1]

$$\begin{array}{l}\min_{\{\boldsymbol{u}_k,\boldsymbol{u}_{k+1},\ldots,\boldsymbol{u}_{k+H_c-1}\}} J \\ \text{with: } J = \sum_{j=1}^{H_p} \|\boldsymbol{e}^y_{k+j}\|^2_{\boldsymbol{Q}_j} + \sum_{j=0}^{H_c-1} \|\Delta\boldsymbol{u}_{k+j}\|^2_{\boldsymbol{S}_j} + \sum_{j=0}^{H_c-1} \|\boldsymbol{e}^u_{k+j}\|^2_{\boldsymbol{R}_j}\end{array} \tag{16.4}$$

subject to the nonlinear model constraints

$$\begin{array}{ll}\boldsymbol{x}_{k+j} = \boldsymbol{f}(\boldsymbol{x}_{k+j-1}, \boldsymbol{u}_{k+j-1}) & \forall\, j = 1, H_p \\ \boldsymbol{y}_{k+j} = g(\boldsymbol{x}_{k+j}) + \boldsymbol{b}_k & \forall\, j = 1, H_p\end{array} \tag{16.5}$$

and the manipulated and controlled variables inequality constraints

$$\begin{array}{ll}\boldsymbol{u}_{min} \le \boldsymbol{u}_{k+j} \le \boldsymbol{u}_{max} & \forall\, j = 0, H_c - 1 \\ \Delta\boldsymbol{u}_{min} \le \Delta\boldsymbol{u}_{k+j} \le \Delta\boldsymbol{u}_{max} & \forall\, j = 0, H_c - 1 \\ \boldsymbol{y}_{j,min} \le \boldsymbol{y}_{k+j} \le \boldsymbol{y}_{j,max} & \forall\, j = 1, H_p\end{array} \tag{16.6}$$

According to model predictive methodology, only the first element $\boldsymbol{u}_k$ of the optimal solution is implemented. $\boldsymbol{e}$ is the deviation between the future output and the reference trajectory. $\boldsymbol{Q}_j$, $\boldsymbol{R}_j$, $\boldsymbol{S}_j$ are semipositive definite matrices which serve as tuning parameters to weight the different contributions in the objective function. Reference trajectories can also be used to avoid violent input variations and implemented in a second optimization.

- Hierarchization of constraints: hard constraints should never be violated, some constraints might be violated and penalized in the objective function.
- Specification of output and input trajectories. The output trajectory can be specified as a set point (which may never be reachable and lead to large input moves), a zone (by soft constraints), a reference trajectory (filter allowing model mismatch and improving robustness as the time constant increases) or a funnel (a kind of min and max reference trajectories)
- Specification of the prediction and control horizons. It is possible to only consider a subset of the prediction horizon (e.g. in the case of nonminimum behaviour) called coincidence points. The prediction horizon should be long enough with respect to the influence of the future manipulated variables' variations on the outputs. The control horizon may be finite or limited to a single move (case of HIECON). In the case of PFC (Adersa), polynomial basis functions are used to parameterize the manipulated variables' profiles.
- Identification. Process input-output data extracted from process tests serve in identification. Signals such as PRBS, or even sometimes simply steps, are currently used. Plant testing is of primordial importance. In linear

[1] $\|\Delta\boldsymbol{u}_k\|^2_S$ is the euclidean norm equal to

$$\|\Delta\boldsymbol{u}_k\|^2_S = \boldsymbol{u}_k \boldsymbol{S} \boldsymbol{u}_k \tag{16.3}$$

identification, in some packages, only one manipulated variable is allowed to vary while the other ones remain at their steady-state value. Some packages (DMI), on the contrary, allow several manipulated variables to vary simultaneously. During the test, the PIDs are fixed; however, operators may intervene to avoid critical situations.

Linear parameter estimation can be performed by two approaches: equation error or output error. Assume a linear state-space model in the form

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}_u\boldsymbol{u}_k + \boldsymbol{B}_v\boldsymbol{v}_k + \boldsymbol{B}_w\boldsymbol{w}_k \\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\xi}_k \end{aligned} \tag{16.7}$$

where $\boldsymbol{v}_k$ are measured disturbances, $\boldsymbol{w}_k$ are unmeasured disturbances or noises and $\boldsymbol{\xi}_k$ are measurement errors. From the state-space representation, a transfer function matrix results

$$\boldsymbol{y}_k = [\boldsymbol{I} - \boldsymbol{\Phi}_y(q^{-1})]^{-1}[\boldsymbol{\Phi}_u(q^{-1})\boldsymbol{u}_k + \boldsymbol{\Phi}_v(q^{-1})\boldsymbol{v}_k + \boldsymbol{\Phi}_w(q^{-1})\boldsymbol{w}_k] + \boldsymbol{\xi}_k \tag{16.8}$$

In the output error identification method, the measurement error $\boldsymbol{\xi}(k)$ is minimized by a nonlinear parameter estimation. The following ARX model results

$$\boldsymbol{y}_k = \boldsymbol{\Phi}_y(q^{-1})\boldsymbol{y}_k + \boldsymbol{\Phi}_u(q^{-1})\boldsymbol{u}_k + \boldsymbol{\Phi}_v(q^{-1})\boldsymbol{v}_k + \boldsymbol{\Phi}_w(q^{-1})\boldsymbol{w}_k + \boldsymbol{\zeta}_k \tag{16.9}$$

with: $\boldsymbol{\zeta}_k = [\boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{y}(q^{-1})\boldsymbol{\xi}_k$.

In the equation error identification method, $\boldsymbol{\zeta}(k)$ is minimized by a linear parameter estimation; however, the noise $\boldsymbol{\zeta}_k$ is coloured even if $\boldsymbol{\xi}_k$ is white.

Discrete step or impulse response data are often used by MPC schemes to represent the process (Li et al., 1989). This makes it easy to use and renders it popular, in particular, in industry. When the system is stable, the transfer function matrix (16.8) is approximated by the following finite impulse response model (used in IDCOM, HIECON, OPC)

$$\boldsymbol{y}_k = \sum_{i=1}^{N_u} \boldsymbol{H}_i^u \boldsymbol{u}_{k-i} + \sum_{i=1}^{N_v} \boldsymbol{H}_i^v \boldsymbol{v}_{k-i} + \sum_{i=1}^{N_w} \boldsymbol{H}_i^w \boldsymbol{w}_{k-i} + \boldsymbol{\xi}_k \tag{16.10}$$

with 30 to 120 coefficients to describe the open-loop response. The model from Eq. (16.10) can be transformed into velocity form by simply replacing each of its variables: y, u, v, w, ξ by its variation Δy, Δu, Δv, Δw, $\Delta\xi$. Alternatively, the following finite step response model can be used (DMC)

$$\begin{aligned} \boldsymbol{y}_k &= \textstyle\sum_{i=1}^{k} \boldsymbol{S}_i^u \Delta\boldsymbol{u}_{k-i} + \sum_{i=1}^{k} \boldsymbol{S}_i^v \Delta\boldsymbol{v}_{k-i} + \sum_{i=1}^{k} \boldsymbol{S}_i^w \Delta\boldsymbol{w}_{k-i} + \boldsymbol{\xi}_k \\ &\text{with: } \boldsymbol{S}_0 = 0, \boldsymbol{S}_i = \boldsymbol{S}_N, \quad \forall i > N \text{ and:} \quad \boldsymbol{H}_i = \boldsymbol{S}_i - \boldsymbol{S}_{i-1} \end{aligned} \tag{16.11}$$

In some cases, continuous Laplace transfer functions are obtained from the discrete-time models, which allows us to later make various transformations according to different sampling periods in discrete time.

All identification methods used in MPC minimize the sum of the squares of errors between measurements and predictions to obtain parameters θ

$$\min_{\theta} \sum_{k=1}^{N_m} \|\hat{y}_k - y_k^m\|^2 \tag{16.12}$$

In equation error identification, the past measurements are fed back to the model (16.9) to obtain the estimates (one-step ahead prediction)

$$\hat{\boldsymbol{y}}_k = \boldsymbol{\Phi}_y(q^{-1})\boldsymbol{y}_k^m + \boldsymbol{\Phi}_u(q^{-1})\boldsymbol{u}_k + \boldsymbol{\Phi}_v(q^{-1})\boldsymbol{v}_k + \boldsymbol{\Phi}_w(q^{-1})\boldsymbol{w}_k \tag{16.13}$$

while in output error identification, the past model estimates are fed back to the model (16.9) to obtain the estimates (several-steps ahead, i.e. long-range prediction)

$$\hat{\boldsymbol{y}}_k = \boldsymbol{\Phi}_y(q^{-1})\hat{\boldsymbol{y}}_k + \boldsymbol{\Phi}_u(q^{-1})\boldsymbol{u}_k + \boldsymbol{\Phi}_v(q^{-1})\boldsymbol{v}_k + \boldsymbol{\Phi}_w(q^{-1})\boldsymbol{w}_k \tag{16.14}$$

Qin and Badgwell (1996) note that FIR models often result in overparameterization and that this problem is overcome by different means such as regularization and partial least squares, in nearly all industrial packages.

- Controller tuning. After the specification of all control design features, after identification and obtaining of a dynamic plant model, any MPC software applied to a given plant needs extensive simulation off-line to test the controller performance. The regulation and tracking are tested as well as the respect of the constraints for controlled and manipulated variables. The robustness must also be addressed by simulation in the case of a plant-model mismatch. After off-line testing, the controller is implemented on-line first in open-loop to verify the model predictions, then in closed-loop where its tuning is improved. DMC uses weights on $\Delta\boldsymbol{u}$ and $\boldsymbol{y}$. IDCOM and HIECON mainly use the time constant of the reference trajectory.

16.2 Linear Model Predictive Control

16.2.1 In the Absence of Constraints

In the absence of constraints, over a finite time horizon, MPC can be seen as GPC (Bitmead et al., 1990; Clarke et al., 1987a,b), which was developed using discrete-time transfer functions and includes a stochastic ARIMAX model. A continuous equivalent of the former discrete predictive control has been developed by Demircioglu and Gawthrop (1991, 1992).

Over an infinite time horizon, linear quadratic (LQ) and linear quadratic Gaussian control (LQG) are formulated in the state-space domain and are inherently multivariable. LQG is a powerful method that is able to handle large-size, nonmimimum-phase systems.

However, the main drawback of the aforementioned control methods is that they do not take into account any type of constraint, including the controlled variables as well as the manipulated variables.

16.2.2 In the Presence of Constraints

Engineers in charge of plants are confronted with constraints regarding the controlled and manipulated variables in particular. Thus, it is not strange that the first successful papers relating MPC dealt with important applications, such as in the petrochemical industry (Cutler and Ramaker, 1979; Richalet et al., 1978).

16.2.3 Short Description of IDCOM

IDCOM belongs to the class of model algorithmic control (MAC) which differs in particular from DMC by its characterization of the process by its impulse response instead of step response for DMC. In IDCOM (Richalet et al., 1978), the process is identified by its impulse responses, where commonly the number of parameters for each impulse response would be $H_c = 40$. Thus, each output $y_j(n)$ of a multivariable system is a weighted sum of the past H_c values $e_k(n-i)$ of the n_u inputs

$$y_j(n) = \sum_{k=1}^{n_u} \sum_{i=1}^{H_c} a_{k,j}(i)\, e_k(n-i) \tag{16.15}$$

with $H_c T_s$ (T_s sampling period) larger than the time response of the system. The identification is not performed continuously and is realized with the system controlled in a supervisory way, which is more favourable for the operators. Pseudo random binary signals are mentioned as appropriate test signals. For given time durations, reference trajectories are defined and the objective of the control algorithm is to compute the future manipulated variables so that the future outputs of the internal model will be as close as possible to this reference trajectory. Constraints are considered on the manipulated variables, their variations and the outputs.

Since the first publication concerning IDCOM, there have been several improvements (IDCOM-M and HIECON: HIErarchical constraint CONtrol). Details of algorithm and industrial applications are cited by Qin and Badgwell (1996). In particular, a multi-objective formulation with quadratic output objective followed by quadratic input objective when the solution to output optimization is not unique, a controllability supervisor deciding which outputs can be independently controlled, computing only one single move for each input which adds robustness at the expense of performance and ranking constraints between hard and soft, are incorporated.

16.2.4 Dynamic Matrix Control (DMC)

Case of a SISO System

The algorithmic principle of MPC is summarized in Figure 16.2: at each time step k, a residual $r(k)$ between the process output $y(k)$ and the output prediction $y^*(k)$ based on past inputs is calculated. Trajectories (i.e. predicted values over the prediction horizon) denoted by $\boldsymbol{y}^*(k)$ and $\boldsymbol{r}^*(k)$ are calculated respectively for $y^*(k)$ and $r(k)$. Then the corrected future output trajectory

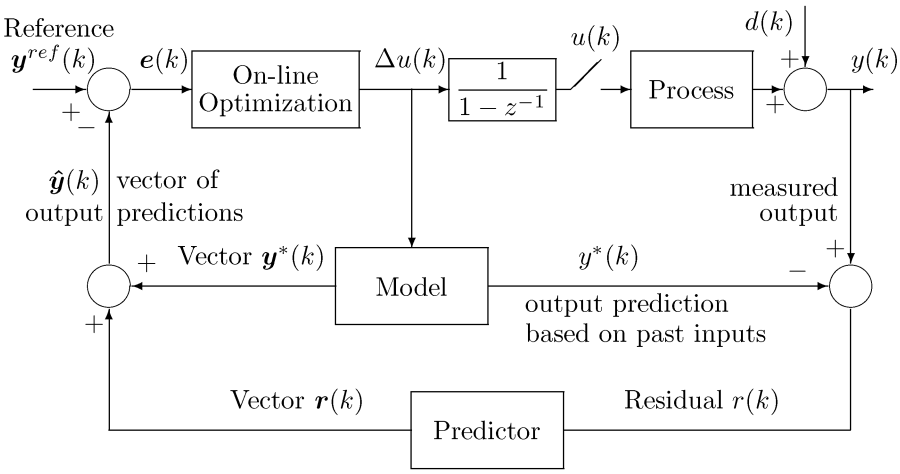


Figure 16.2: General block diagram of Model Predictive Control

$\hat{\boldsymbol{y}}(k)$ is obtained, which is compared to the desired (or reference) trajectory $\boldsymbol{y}^{ref}(k)$ from which the MPC control law results.

A main drawback of DMC is that it can be only used for stable plants and also without integrators: this drawback comes from the fact that only the first M coefficients of the step response are considered for modelling the process in Eq. (16.16). This can be enhanced by separating the DMC algorithm into a predictor and an optimizer (Lunström et al., 1995). This avoids, in particular, using too many step response coefficients as in classical DMC.

Here, classical DMC is first discussed. In DMC (Cutler and Ramaker, 1979; Garcia and Morshedi, 1986), a single-input single-output system is represented by its truncated step response model

$$y(j+1) = y_{ss} + \sum_{i=1}^{M} h_i \, \Delta u(j+1-i) + d(j+1) \tag{16.16}$$

where h_i are the unit step response coefficients for the i-th time interval, y_{ss} is the initial steady-state output (y is not a deviation variable) and $\boldsymbol{d}$ represents the unmodelled factors affecting the outputs. Also, $\Delta u(k) = u(k) - u(k-1)$. M is the number of time intervals necessary to reach the steady state; it will be called the model horizon or truncation number (thus, $h_i = h_M$ if $i \geq M$).

In the case where the coefficients $\bar{g}_i$ of the unit impulse response are used, the truncated impulse response model results

$$y(j+1) = y_{ss} + \sum_{i=1}^{M} \bar{g}_i \, u(j+1-i) + d(j+1) \tag{16.17}$$

The unit step response coefficients h_i are related to the unit impulse coefficients

$\bar{g}_i$ by the relations

$$\bar{g}_i = h_i - h_{i-1} \qquad \text{and:} \qquad h_i = \sum_{j=1}^{i} \bar{g}_j \tag{16.18}$$

Rigorously, if M was the truncation order of the step response model, the truncated step response model (Garcia et al., 1989) should be written as

$$y(j+1) = y_{ss} + \sum_{i=1}^{M-1} h_i \, \Delta u(j+1-i) + h_M \left(u(j+1-M) - u_{ss}\right) \tag{16.19}$$

but the last term is often omitted on purpose, however for set point tracking it must be taken into account. In this expression, u is considered as an absolute variable and not a deviation variable hence the use of its initial or steady-state value u_{ss}. At time k, $y(k)$ is known and $u(k)$ is to be determined.

Considering a prediction horizon H_p and the set point y^s, the objective is to compute the future inputs so that the future outputs will be close to the set point. Thus, at time $k+l$, the output prediction based on past and future inputs is decomposed into

$$\hat{y}(k+l|k) = \underbrace{y_{ss} + \sum_{i=l+1}^{M-1} h_i \, \Delta u(k+l-i) + h_M \left(u(k+l-M) - u_{ss}\right)}_{\text{effect of past inputs}} + \underbrace{\sum_{i=1}^{l} h_i \, \Delta u(k+l-i)}_{\text{effect of future inputs}} + \underbrace{\hat{d}(k+l|k)}_{\text{effect of predicted disturbances}} \tag{16.20}$$

At each instant k, only H_c future input changes are computed, so that

$$\Delta u(j) = 0 \qquad \forall \; j \geq k + H_c \tag{16.21}$$

Thus, beyond the control horizon H_c, i.e. after instant $k+H_c$, the manipulated input is assumed to be constant.

From Eq. (16.20), define $y^*(k+l|k)$ as the output prediction corresponding to the influence of the past input variations equal to

$$y^*(k+l|k) = y_{ss} + \sum_{i=l+1}^{M-1} h_i \, \Delta u(k+l-i) + h_M \left(u(k+l-M) - u_{ss}\right) \tag{16.22}$$

If $l \geq M-1$, Eq. (16.23) is simplified as

$$y^*(k+l|k) = y_{ss} + h_M \left(u(k-1) - u_{ss}\right) \tag{16.23}$$

and Eq. (16.20) is reduced to

$$\hat{y}(k+l|k) = y_{ss} + h_M \left(u(k-1) - u_{ss}\right) + \sum_{i=l-H_c+1}^{l} h_i \, \Delta u(k+l-i) + \hat{d}(k+l|k) \tag{16.24}$$

Over a given prediction horizon H_p, and assuming $M > H_c$, the vector of output predictions can be decomposed into

$$\begin{aligned}
\hat{y}(k+1|k) &= h_1\Delta u(k) \\
&\quad +y_{ss} + h_2\Delta u(k-1) + \ldots + h_{M-1}\Delta u(k-M+2) \\
&\quad +h_M\left(u(k+1-M) - u_{ss}\right) + \hat{d}(k+1|k) \\
&= h_1\Delta u(k) + y^*(k+1|k) + \hat{d}(k+1|k) \\
\hat{y}(k+2|k) &= h_1\Delta u(k+1) + h_2\Delta u(k) \\
&\quad +y_{ss} + h_3\Delta u(k-1) + \ldots + h_{M-1}\Delta u(k-M+3) \\
&\quad +h_M\left(u(k+2-M) - u_{ss}\right) + \hat{d}(k+2|k) \\
&= h_1\Delta u(k+1) + h_2\Delta u(k) + y^*(k+2|k) + \hat{d}(k+2|k) \\
&\vdots \\
\hat{y}(k+H_c|k) &= h_1\Delta u(k+H_c-1) + h_2\Delta u(k+H_c-2) + \ldots + h_{H_c}\Delta u(k) \\
&\quad +y_{ss} + h_{H_c+1}\Delta u(k-1) + \ldots + h_{M-1}\Delta u(k+H_c-M+1) \\
&\quad +h_M\left(u(k+H_c-M) - u_{ss}\right) + \hat{d}(k+H_c|k) \\
&= h_1\Delta u(k+H_c-1) + \ldots + h_{H_c}\Delta u(k) + y^*(k+H_c|k) + \hat{d}(k+H_c|k)
\end{aligned}$$

$$\begin{aligned}
\hat{y}(k+H_c+1|k) &= h_2\Delta u(k+H_c-1) + \ldots + h_{H_c+1}\Delta u(k) \\
&\quad +y^*(k+H_c+1|k) + \hat{d}(k+H_c+1|k) \\
&\vdots \\
\hat{y}(k+M|k) &= h_{M-H_c+1}\Delta u(k+H_c-1) + \ldots + h_M\Delta u(k) \\
&\quad +y^*(k+M|k) + \hat{d}(k+M|k) \\
\hat{y}(k+M+1|k) &= h_{M-H_c+2}\Delta u(k+H_c-1) + \ldots + h_M\Delta u(k+1) + h_M\Delta u(k) \\
&\quad +y^*(k+M+1|k) + \hat{d}(k+M+1|k) \\
&\vdots \\
\hat{y}(k+H_p|k) &= h_M\Delta u(k+H_c-1) + \ldots + h_M\Delta u(k) \\
&\quad +y^*(k+H_p|k) + \hat{d}(k+H_p|k) \\
&\qquad \text{if } H_p \geq H_c + M - 1
\end{aligned} \tag{16.25}$$

resulting in

$$\begin{bmatrix} \hat{y}(k+1|k) \\ \vdots \\ \hat{y}(k+H_p|k) \end{bmatrix} = \begin{bmatrix} y^*(k+1|k) \\ \vdots \\ y^*(k+H_p|k) \end{bmatrix} + \boldsymbol{\mathcal{A}} \begin{bmatrix} \Delta u(k) \\ \vdots \\ \Delta u(k+H_c-1) \end{bmatrix} + \begin{bmatrix} \hat{d}(k+1|k) \\ \vdots \\ \hat{d}(k+H_p|k) \end{bmatrix} \tag{16.26}$$

where $\boldsymbol{\mathcal{A}}$ is the $H_p \times H_c$ dynamic matrix (hence dynamic matrix control) of the

system equal to

$$\boldsymbol{\mathcal{A}} = \begin{bmatrix} h_1 & 0 & \dots & 0 \\ h_2 & h_1 & & \vdots \\ \vdots & \vdots & \ddots & \\ h_{H_c} & h_{H_c-1} & \dots & h_1 \\ \vdots & \vdots & & \vdots \\ h_M & h_{M-1} & \dots & h_{M-H_c+1} \\ \vdots & \vdots & & \vdots \\ h_M & h_M & \dots & h_M \\ \vdots & \vdots & & \vdots \\ h_M & h_M & \dots & h_M \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \left.\begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix}\right\} H_c \text{ rows} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} (M-1) \text{ rows} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} (H_p - H_c - M + 1) \text{ rows} \tag{16.27}$$

Equation (16.26) thus shows the influence of future inputs. The dynamic matrix $\boldsymbol{\mathcal{A}}$ of Eq. (16.27) is exactly similar to the matrix $\boldsymbol{G}$ of Eq. (15.23) defined in GPC.

The vector of output predictions $y^*(k+l|k)$ corresponding to the influence of the past input variations can itself be calculated (assuming $H_p \geq M$) as

$$\begin{bmatrix} y^*(k+1|k) \\ \vdots \\ y^*(k+M-1|k) \\ y^*(k+M|k) \\ \vdots \\ y^*(k+H_p|k) \end{bmatrix} = \begin{bmatrix} y_{ss} \\ \vdots \\ y_{ss} \end{bmatrix} + \begin{bmatrix} h_{M-1} & h_{M-2} & \dots & h_2 \\ 0 & h_{M-1} & & \vdots \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & h_{M-1} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \Delta u(k-M+2) \\ \vdots \\ \Delta u(k-1) \end{bmatrix}$$
$$+ h_M \begin{bmatrix} u(k-M+1) - u_{ss} \\ \vdots \\ u(k-1) - u_{ss} \\ u(k-1) - u_{ss} \\ \vdots \\ u(k-1) - u_{ss} \end{bmatrix} \tag{16.28}$$

The combination of Eqs. (16.16) for $j = k-1$ and (16.20) for $l = 0$ gives the influence of the unmodelled effects

$$d(k) = y(k) - y^*(k|k) \tag{16.29}$$

Consequently, based on a measured output $y^m(k)$, an estimation of $d(k)$ is

given as

$$\begin{aligned}\hat{d}(k+l|k) &= \hat{d}(k|k) = y^m(k) - y^*(k|k) \qquad \forall\, l = 1, \ldots, H_p \\ &= y^m(k) - [y_{ss} + \sum_{i=1}^{M-1} h_i \Delta u(k-i) + h_M(u(k-M) - u_{ss})]\end{aligned} \tag{16.30}$$

thus the predicted disturbances are all equal to the present estimated disturbance.

Let us define a quadratic criterion, taking into account the difference between the estimated output and the reference over the time horizon as

$$J = \sum_{i=1}^{H_p} (\hat{y}(k+i|k) - y^{\text{ref}}(k+i))^2 \tag{16.31}$$

From the previous equations, this can be seen as computing the vector of future input variations

$$\Delta \boldsymbol{u}(k) = [\Delta u(k) \ldots \Delta u(k + H_c - 1)]^T \tag{16.32}$$

which is the least-squares solution of the following linear system resulting from Eq. (16.26)

$$\begin{bmatrix} y^{\text{ref}}(k+1) - y^*(k+1|k) - \hat{d}(k|k) = e(k+1) \\ \vdots \\ y^{\text{ref}}(k+H_p) - y^*(k+H_p|k) - \hat{d}(k|k) = e(k+H_p) \end{bmatrix} = \boldsymbol{e}(k+1) = \boldsymbol{\mathcal{A}}\, \Delta \boldsymbol{u}(k) \tag{16.33}$$

The least-squares solution of (16.33) is

$$\Delta \boldsymbol{u}(k) = (\boldsymbol{\mathcal{A}}^T \boldsymbol{\mathcal{A}})^{-1} \boldsymbol{\mathcal{A}}^T\, \boldsymbol{e}(k+1) \tag{16.34}$$

Only the first input variation of vector (16.32) equal to (16.34) is implemented. According to Garcia and Morshedi (1986), the choice of the time horizon H_p as $H_p = H_c + M$ generally results in a stable controller. Garcia et al. (1989) simply state that for sufficiently small H_c and sufficiently large H_p such that $H_p > H_c + M - 1$, the closed-loop system is stable. The control law (16.34) provides a too strong control action and will be improved in the following by introduction of weighting terms as in (16.45). However, Li et al. (1989) assume $M \geq H_p$ and Camacho and Bordons (1998) choose $M \gg H_p$. Soeterboek (1992) shows that the prediction horizon must be larger when constraints are present.

Case of a MIMO system

Similarly, a multivariable system (n_u inputs, n_y outputs) is represented as

$$\boldsymbol{y}(k+1) = \boldsymbol{y}_0 + \sum_{i=1}^{M} \boldsymbol{a}_i\, \Delta \boldsymbol{u}(k-i+1) + \boldsymbol{d}(k+1) \tag{16.35}$$

where $\boldsymbol{a}_i$ is an $n_y \times n_u$ matrix of unit step response coefficients for the i-th time interval, $\boldsymbol{y}_0$ is the initial output vector and $\boldsymbol{d}$ represents the unmodelled factors affecting the outputs.

Any input-output pair i-j can be represented by a matrix $\boldsymbol{\mathcal{A}}_{ij}$ of coefficients h perfectly similar to Eq. (16.27) so that the complete system is finally represented by a multivariable dynamic matrix composed of elementary matrices $\boldsymbol{\mathcal{A}}_{ij}$ of type (16.27) as

$$\boldsymbol{\mathcal{A}} = \begin{bmatrix} \boldsymbol{\mathcal{A}}_{11} & \dots & \boldsymbol{\mathcal{A}}_{1n_u} \\ \vdots & & \vdots \\ \boldsymbol{\mathcal{A}}_{n_y1} & \dots & \boldsymbol{\mathcal{A}}_{n_yn_u} \end{bmatrix} \tag{16.36}$$

For a 2×2 system, the representation of the system equivalent to Eq. (16.26) would be

$$\begin{aligned} \begin{bmatrix} \hat{y}_1(k+1|k) \\ \vdots \\ \hat{y}_1(k+H_p|k) \end{bmatrix} &= \begin{bmatrix} y_1^*(k+1|k) \\ \vdots \\ y_1^*(k+H_p|k) \end{bmatrix} + \begin{bmatrix} \hat{d}_1(k+1|k) \\ \vdots \\ \hat{d}_1(k+H_p|k) \end{bmatrix} \\ &+\boldsymbol{\mathcal{A}}_{11} \begin{bmatrix} \Delta u_1(k) \\ \vdots \\ \Delta u_1(k+H_c-1) \end{bmatrix} + \boldsymbol{\mathcal{A}}_{12} \begin{bmatrix} \Delta u_2(k) \\ \vdots \\ \Delta u_2(k+H_c-1) \end{bmatrix} \\ \begin{bmatrix} \hat{y}_2(k+1|k) \\ \vdots \\ \hat{y}_2(k+H_p|k) \end{bmatrix} &= \begin{bmatrix} y_2^*(k+1|k) \\ \vdots \\ y_2^*(k+H_p|k) \end{bmatrix} + \begin{bmatrix} \hat{d}_2(k+1|k) \\ \vdots \\ \hat{d}_2(k+H_p|k) \end{bmatrix} \\ &+\boldsymbol{\mathcal{A}}_{21} \begin{bmatrix} \Delta u_1(k) \\ \vdots \\ \Delta u_1(k+H_c-1) \end{bmatrix} + \boldsymbol{\mathcal{A}}_{22} \begin{bmatrix} \Delta u_2(k) \\ \vdots \\ \Delta u_2(k+H_c-1) \end{bmatrix} \end{aligned} \tag{16.37}$$

or, finally,

$$\begin{aligned} \begin{bmatrix} \hat{y}_1(k+1|k) \\ \vdots \\ \hat{y}_1(k+H_p|k) \\ \hat{y}_2(k+1|k) \\ \vdots \\ \hat{y}_2(k+H_p|k) \end{bmatrix} &= \begin{bmatrix} y_1^*(k+1|k) \\ \vdots \\ y_1^*(k+H_p|k) \\ y_2^*(k+1|k) \\ \vdots \\ y_2^*(k+H_p|k) \end{bmatrix} + \begin{bmatrix} \hat{d}_1(k+1|k) \\ \vdots \\ \hat{d}_1(k+H_p|k) \\ \hat{d}_2(k+1|k) \\ \vdots \\ \hat{d}_2(k+H_p|k) \end{bmatrix} \\ &+ \begin{bmatrix} \boldsymbol{\mathcal{A}}_{11} & \boldsymbol{\mathcal{A}}_{12} \\ \boldsymbol{\mathcal{A}}_{21} & \boldsymbol{\mathcal{A}}_{22} \end{bmatrix} \begin{bmatrix} \Delta u_1(k) \\ \vdots \\ \Delta u_1(k+H_c-1) \\ \Delta u_2(k) \\ \vdots \\ \Delta u_2(k+H_c-1) \end{bmatrix} \end{aligned} \tag{16.38}$$

Define the vector of future input variations similarly to (16.32)

$$\Delta \boldsymbol{u}(k) = [\Delta \boldsymbol{u}_1(k)^T \ \ldots \ \Delta \boldsymbol{u}_{n_u}(k)^T]^T \tag{16.39}$$

and also the vector of deviations similarly to (16.33)

$$\boldsymbol{e}(k+1) = [\boldsymbol{e}_1(k+1)^T \ \ldots \ \boldsymbol{e}_{n_y}(k+1)^T]^T \tag{16.40}$$

The least-squares solution of the multivariable DMC controller is also given by (16.34). Again, Garcia and Morshedi (1986) recommend the choice $H_p = H_c + M$ for obtaining a stable controller. A large value of the prediction horizon H_p improves stability even if it does not significantly improve performance (Shridar and Cooper, 1998). The control horizon H_c should be taken to be greater than 1.

Some input variations can be suppressed by formulating multivariable DMC as

$$\begin{bmatrix} \boldsymbol{e}(k+1) \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mathcal{A}} \\ \boldsymbol{\Lambda} \end{bmatrix} \Delta \boldsymbol{u}(k) \tag{16.41}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix equal to

$$\boldsymbol{\Lambda} = \text{diag}(\underbrace{\lambda_1 \ldots \lambda_1}_{H_c \text{ values}} \ \lambda_2 \ldots \lambda_2 \ldots \ldots \lambda_{n_u} \ldots \lambda_{n_u}) \tag{16.42}$$

It is also possible to selectively weight the controlled variables by multiplying the equations such as (16.33). The matrix of weights is thus

$$\boldsymbol{\Gamma} = \text{diag}(\underbrace{\gamma_1 \ldots \gamma_1}_{H_p \text{ values}} \ \gamma_2 \ldots \gamma_2 \ldots \ldots \gamma_{n_y} \ldots \gamma_{n_y}) \tag{16.43}$$

The matrix $\boldsymbol{\Gamma}$ defined in (16.43) for selective weighting of controlled variables and the matrix $\boldsymbol{\Lambda}$ defined in (16.42) for suppression of some input variations are now incorporated in the criterion. The following quadratic criterion to be minimized with respect to $\Delta \boldsymbol{u}(k)$ results

$$\begin{aligned} J =& \frac{1}{2} \left[\hat{\boldsymbol{y}}(k) - \boldsymbol{y}^{ref}(k) \right]^T \boldsymbol{\Gamma}^T \boldsymbol{\Gamma} \left[\hat{\boldsymbol{y}}(k) - \boldsymbol{y}^{ref}(k) \right] + \frac{1}{2} \Delta \boldsymbol{u}^T(k) \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \Delta \boldsymbol{u}(k) \\ =& \frac{1}{2} \left[\boldsymbol{\mathcal{A}} \Delta \boldsymbol{u}(k) - \boldsymbol{e}(k+1) \right]^T \boldsymbol{\Gamma}^T \boldsymbol{\Gamma} \left[\boldsymbol{\mathcal{A}} \Delta \boldsymbol{u}(k) - \boldsymbol{e}(k+1) \right] + \frac{1}{2} \Delta \boldsymbol{u}^T(k) \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \Delta \boldsymbol{u}(k) \end{aligned} \tag{16.44}$$

In the absence of constraints, the solution of (16.44) is

$$\Delta \boldsymbol{u}(k) = \left(\boldsymbol{\mathcal{A}}^T \boldsymbol{\Gamma}^T \boldsymbol{\Gamma} \boldsymbol{\mathcal{A}} + \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \right)^{-1} \boldsymbol{\mathcal{A}}^T \boldsymbol{\Gamma}^T \boldsymbol{\Gamma} \boldsymbol{e}(k+1) \tag{16.45}$$

Again, the control law of DMC in the absence of constraints given by Eq. (16.45) is similar to the control law of GPC given by Eq. (15.26).

A typical industrial study concerning application of DMC to a model IV fluid catalytic cracker is related by Gusciora et al. (1992) who describe the linear

matrix model used for control with 11 inputs and 9 outputs, the DMC block diagram, the main problems encountered and the benefits generated . Shridar and Cooper (1998) propose a tuning strategy for unconstrained multivariable DMC. They note that the prediction horizon H_p should be in accordance with the settling time of the process. Also, increasing the input horizon H_c from 2 to 6 does not modify greatly the closed-loop performance, but it should be greater than or equal to the number of unstable modes of the system (Rawlings and Muske, 1993). Al-Ghazzawi et al. (2001) present an on-line tuning strategy based on the use of sensitivity functions for the closed-loop response with respect to the MPC tuning parameters.

16.2.5 Quadratic Dynamic Matrix Control (QDMC)

The handling of constraints was not completely satisfactory in the original DMC which motivated Garcia and Morshedi (1986) to develop a quadratic programming solution to the DMC problem. Different types of constraints (soft, which can be violated, and hard constraints) are commonly encountered:

- Constraints affecting the manipulated variables such as valve saturations

$$\boldsymbol{u}_{\min} \leq \boldsymbol{u} \leq \boldsymbol{u}_{\max} \tag{16.46}$$

- Constraints affecting the controlled variables: e.g. avoid overshoots.
- Constraints affecting other variables which must be kept within bounds.
- Constraints added to the process in order to avoid inverse responses resulting in nonminimum-phase behaviour.
- Terminal state constraints.

All these constraints can be resumed as a system of linear inequalities incorporating the dynamic information concerning the projection of constraints

$$\boldsymbol{B}\,\Delta\boldsymbol{u}(k) \leq \boldsymbol{c}(k+1) \tag{16.47}$$

where $\boldsymbol{B}$ contains dynamic information on the constraints and $\boldsymbol{c}(k+1)$ contains projected deviations of constrained variables and their bounds. It has been stated in the literature (Camacho and Bordons, 1998; Garcia and Morshedi, 1986; Maciejowski, 2002; Soeterboek, 1992) how to represent the aforementioned cases of constraints; note that $\Delta\boldsymbol{u}(k)$ also contains the prediction of the future input changes.

In the presence of constraints (16.46) and (16.47), the problem can be thus formulated as quadratic programming, such as

$$\min_{\Delta\boldsymbol{u}(k)} \left[\frac{1}{2}\Delta\boldsymbol{u}(k)^T\boldsymbol{H}\Delta\boldsymbol{u}(k) - \boldsymbol{g}(k+1)^T\Delta\boldsymbol{u}(k)\right] \tag{16.48}$$

subject to constraints (16.46) and (16.47). $\boldsymbol{H}$ is the Hessian matrix (in general fixed) equal to

$$\boldsymbol{H} = \boldsymbol{\mathcal{A}}^T\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}\boldsymbol{\mathcal{A}} + \boldsymbol{\Lambda}^T\boldsymbol{\Lambda} \tag{16.49}$$

and $\boldsymbol{g}(k+1)$ is the gradient vector equal to

$$\boldsymbol{g}(k+1) = \boldsymbol{\mathcal{A}}^T \boldsymbol{\Gamma}^T \boldsymbol{\Gamma} \boldsymbol{e}(k+1) \tag{16.50}$$

This quadratic problem can be solved efficiently by available numerical subroutines based, for example, on Rosen's method (Soeterboek, 1992), conjugate gradients or a quasi-Newton method (Camacho and Bordons, 1998; Fletcher, 1991).

In QDMC, the choice of the projection interval to be constrained is important, as not all the horizon H_p necessarily needs to be constrained. For example, for a nonminimum-phase system (dead time or inverse response), shifting the constraint window towards the end of the horizon is favourable.

Another version of DMC, called LDMC, where the criterion concerns the sum of the absolute values of the errors has been developed (Morshedi et al., 1985). In this case, the optimization problem is solved by linear programming.

In some cases, stability due to output constraints can be found with QDMC (Muske and Rawlings, 1993) for open-loop unstable systems.

16.2.6 State-Space Formulation of Dynamic Matrix Control

Assume that the system is described by a state-space discrete-time model of the form

$$\begin{array}{ll} \boldsymbol{x}(k+1) & = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\boldsymbol{u}(k) \\ \boldsymbol{y}(k) & = \boldsymbol{C}\boldsymbol{x}(k) \end{array} \tag{16.51}$$

Assuming zero-initial conditions (x, u and y are deviation variables), the equivalent discrete-time transfer matrix is

$$\bar{G}(z) = \frac{Y(z)}{U(z)} = \boldsymbol{C}(z\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B} \tag{16.52}$$

Assuming that $\boldsymbol{A}$ is stable, $\bar{G}(z)$ can be expanded as

$$\bar{G}(z) = \sum_{i=1}^{\infty} \bar{g}_i z^{-i} \qquad \Longleftrightarrow Y(z) = \sum_{i=1}^{\infty} \bar{g}_i z^{-i} U(z) \tag{16.53}$$

where $\bar{g}_i$ are coefficients of the impulse response, decreasing exponentially. For this reason, in the time domain, the impulse response is generally truncated at order n as

$$y(k) = \sum_{i=1}^{n} \bar{g}_i u(k-i) \tag{16.54}$$

If the input u is chosen as a unit step, it results

$$y(0) = 0\ ,\ y(1) = \bar{g}_1\ ,\ y(2) = \bar{g}_1 + \bar{g}_2\ ,\ \ldots\ ,\ y(k) = \sum_{i=1}^{k} \bar{g}_i \tag{16.55}$$

Define

$$h_k = \sum_{i=1}^{k} \bar{g}_i \Longrightarrow y(k) = h_k \tag{16.56}$$

The h_k are the step response coefficients. The following relation can also be given

$$\bar{g}_i = h_i - h_{i-1} \tag{16.57}$$

Equation (16.54) can be expressed with respect to the step response coefficients as

$$y(k) = \sum_{i=1}^{n-1} h_i \Delta u(k-i) + h_n u(k-n) \qquad \text{with:} \quad \Delta u(k) = u(k) - u(k-1) \tag{16.58}$$

Thus, any state-space discrete-time system can be transformed into a truncated step response model which can be later used in the DMC or QDMC formulation (Garcia et al., 1989). However, this transformation remains artificial and does not constitute a real state-space method.

16.2.7 State-Space Linear Model Predictive Control as OBMPC

The state-space approach is appealing, as it allows us to avoid problems due to the truncation of the step response in the classical DMC approach. Li et al. (1989) proposed first state-space formulation for MPC. State-space linear MPC was developed in a similar spirit (Lee et al., 1994; Lunström et al., 1995; Ricker, 1990a). In these papers, two stages are used for the model prediction: first, a predictor of the output is built, then a state observer is used to estimate the states. The observer-based model predictive control (OBMPC) developed by Lee et al. (1994), taken again by Lunström et al. (1995), is discussed in the following.

The original idea (Li et al., 1989) is to represent the entire trajectory for a SISO system as a sequence of states

$$\begin{aligned} &x_1(k) = x_2(k-1) + h_1\, \Delta u(k-1) \\ &\dots \\ &x_i(k) = x_{i+1}(k-1) + h_i\, \Delta u(k-1) \qquad i = 1, M \end{aligned} \tag{16.59}$$

where h_i are the step response coefficients, and to use these equations with respect to the predicted output corresponding to the influence of past variations

$$y^*(k+i-1|k) = y^*(k+i-1|k-1) + h_i\, \Delta u(k-1) \tag{16.60}$$

Note that the influence of the input disappears for a stable process when $i > M$ as $h_i = h_{ss}$ (steady-state).

In the case of a MIMO system with n_u inputs and n_y outputs, the matrix $\boldsymbol{S}_i$ is defined at each instant i to represent each coefficient of the step responses

as

$$\boldsymbol{S}_i = \begin{bmatrix} h_{1,1,i} & h_{1,2,i} & \dots & h_{1,n_u,i} \\ h_{2,1,i} & h_{2,2,i} & \dots & h_{2,n_u,i} \\ \vdots & \vdots & \dots & \vdots \\ h_{n_y,1,i} & h_{n_y,2,i} & \dots & h_{n_y,n_u,i} \end{bmatrix} \tag{16.61}$$

where $h_{k,l,i}$ is the step coefficient at instant i of output k with step input l.

It is assumed that at time k, the inputs to be determined are $u(k)$ and the future control actions.

In the absence of disturbances, the state-space form corresponding to the step response model can then be written as

$$\begin{array}{ll} \boldsymbol{Y}(k) & = \boldsymbol{\Phi}\,\boldsymbol{Y}(k-1) + \boldsymbol{S}\,\Delta\boldsymbol{u}(k-1) \\ \boldsymbol{y}^*(k|k) & = \boldsymbol{\Psi}\,\boldsymbol{Y}(k) \end{array} \tag{16.62}$$

This original state-space model (16.62) can be extended (Lunström et al., 1995) in a similar way to the LQG state-space model to include disturbances w and output noise v as

$$\begin{array}{ll} \boldsymbol{Y}(k) & = \boldsymbol{\Phi}\,\boldsymbol{Y}(k-1) + \boldsymbol{S}\,\Delta\boldsymbol{u}(k-1) + \boldsymbol{T}\,\Delta\boldsymbol{w}(k-1) \\ \boldsymbol{y}^*(k|k) & = \boldsymbol{\Psi}\,\boldsymbol{Y}(k) \\ \boldsymbol{y}(k) & = \boldsymbol{y}^*(k|k) + \boldsymbol{v}(k) \end{array} \tag{16.63}$$

with

$$\begin{array}{l} \boldsymbol{Y}(k) = [\boldsymbol{y}^*(k|k)^T \;\; \boldsymbol{y}^*(k+1|k)^T \;\; \dots \boldsymbol{y}^*(k+M-1|k)^T \; \boldsymbol{x}_u(k)^T \; \boldsymbol{x}_w(k)^T]^T \\ \boldsymbol{y}^*(k+i|k) = [y_1^*(k+i|k) \dots y_{n_y}^*(k+i|k)]^T \\ \Delta\boldsymbol{u}(k) = [\Delta u_1(k) \dots \Delta u_{n_u}(k)]^T \end{array} \tag{16.64}$$

the matrix $\boldsymbol{\Phi}$ of size $(M.n_y + \dim \boldsymbol{x}_u + \dim \boldsymbol{x}_w)$

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & \boldsymbol{I}_{n_y} & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{I}_{n_y} & \ddots & 0 & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \dots & & \\ 0 & 0 & \dots & & \boldsymbol{I}_{n_y} & 0 & 0 \\ 0 & 0 & \dots & & \boldsymbol{I}_{n_y} & \boldsymbol{C}_u & \boldsymbol{C}_w \\ 0 & 0 & \dots & & 0 & \boldsymbol{A}_u & 0 \\ 0 & 0 & \dots & & 0 & 0 & \boldsymbol{A}_w \end{bmatrix} \tag{16.65}$$

the matrix $\boldsymbol{S}$ equal to

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_1 \\ \vdots \\ \boldsymbol{S}_M \\ \boldsymbol{B}_u \\ 0 \end{bmatrix} \tag{16.66}$$

the matrix $\boldsymbol{\Psi}$ equal to

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{I}_{n_y} & \boldsymbol{0} & \dots & \boldsymbol{0} \end{bmatrix} \tag{16.67}$$

and the matrix $\boldsymbol{T}$ equal to

$$\boldsymbol{T} = \left[\begin{array}{cccc} \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{B}_w \end{array} \right]^T \tag{16.68}$$

Residual plant dynamics and disturbance dynamics have been respectively incorporated in (16.63) in state-space form as

$$\begin{array}{l} \boldsymbol{x}_u(k+1) = \boldsymbol{A}_u\, \boldsymbol{x}_u(k) + \boldsymbol{B}_u\, \Delta \boldsymbol{u}(k) \\ \boldsymbol{x}_w(k+1) = \boldsymbol{A}_w\, \boldsymbol{x}_d(k) + \boldsymbol{B}_w\, \Delta \boldsymbol{w}(k) \end{array} \tag{16.69}$$

Moreover, the deviation between the last and previous predictions considers the additional residual plant and disturbance dynamics as

$$\boldsymbol{y}^*(k+M-1|k) = \boldsymbol{y}^*(k+M-1|k-1) + \boldsymbol{S}_M\, \Delta \boldsymbol{u}(k) + \boldsymbol{C}_u\, \boldsymbol{x}_u(k) + \boldsymbol{C}_w\, \boldsymbol{x}_w(k) \tag{16.70}$$

The future outputs are predicted by using a state observer such as the optimal linear Kalman filter of matrix gain $\boldsymbol{K}$. In a general way, the predictions of the future outputs would be given by the two-stage form: first, the model prediction

$$\hat{\boldsymbol{Y}}(k+1|k) = \boldsymbol{\Phi}\, \hat{\boldsymbol{Y}}(k|k) + \boldsymbol{S}\, \Delta \boldsymbol{u}(k) \tag{16.71}$$

then, the correction based on measurements

$$\hat{\boldsymbol{Y}}(k|k) = \hat{\boldsymbol{Y}}(k|k-1) + \boldsymbol{K}\, [\boldsymbol{y}(k) - \hat{\boldsymbol{y}}^*(k|k-1)] \tag{16.72}$$

with

$$\begin{array}{lcl} \hat{\boldsymbol{Y}}(k|k-1) & = & [\hat{\boldsymbol{y}}^*(k|k-1)^T \;\; \hat{\boldsymbol{y}}^*(k+1|k-1)^T \;\; \ldots \hat{\boldsymbol{y}}^*(k+M-1|k)^T \\ & & \hat{\boldsymbol{x}}_u \;\; \hat{\boldsymbol{x}}_w]^T \\ \hat{\boldsymbol{y}}^*(k|k-1) & = & \boldsymbol{\Psi}\, \hat{\boldsymbol{Y}}(k|k) \end{array} \tag{16.73}$$

Now that the model states are estimated, the optimization can be performed in a similar way to QDMC (Garcia and Morshedi, 1986). The objective function to be minimized with respect to $\Delta \boldsymbol{\mathcal{U}}(k)$ is

$$J = \| \, \Gamma \, [\mathcal{Y}(k+1|k) - \mathcal{R}(k+1|k)] \, \|^2 + \| \, \Lambda \, \Delta \mathcal{U}(k|k) \, \|^2 \tag{16.74}$$

with

$$\begin{array}{ll} \Delta \mathcal{U}(k|k) & = [\, \Delta \boldsymbol{u}(k|k)^T \, \Delta \boldsymbol{u}(k+1|k)^T \ldots \Delta \boldsymbol{u}(k+H_c-1|k)^T \,]^T \\ \mathcal{Y}(k+1|k) & = \boldsymbol{\Phi}_{H_p}\, \hat{\boldsymbol{Y}}(k|k) + \boldsymbol{\mathcal{S}}_{H_p}\, \Delta \mathcal{U}(k|k) \\ & = [\, \boldsymbol{y}_f(k+1|k)^T \; \ldots \; \boldsymbol{y}_f(k+H_p|k)^T \,]^T \\ \mathcal{R}(k+1|k) & = [\, \boldsymbol{r}(k+1|k)^T \; \ldots \; \boldsymbol{r}(k+H_p|k)^T \,]^T \end{array} \tag{16.75}$$

where $\mathcal{R}$ is the reference trajectory. Note that the H_c future input moves are used in $\mathcal{Y}(k+1|k)$ of components $\boldsymbol{y}_f$. The matrix $\boldsymbol{\mathcal{S}}_{H_p}$ is

$$\boldsymbol{\mathcal{S}}_{H_p} = \left[\begin{array}{ccccc} \boldsymbol{S}_1 & 0 & 0 & \ldots & 0 \\ \boldsymbol{S}_2 & \boldsymbol{S}_1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \boldsymbol{S}_{H_c} & \boldsymbol{S}_{H_c-1} & \ldots & & \boldsymbol{S}_1 \\ \vdots & \vdots & & & \vdots \\ \boldsymbol{S}_{H_p} & \boldsymbol{S}_{H_p-1} & \ldots & & \boldsymbol{S}_{H_p-H_c+1} \end{array} \right] \tag{16.76}$$

and $\boldsymbol{\Phi}_{H_p}$ is

$$\boldsymbol{\Phi}_{H_p} = \left[\begin{array}{cc} \boldsymbol{I}_{n_y.H_p} & \mathbf{0} \end{array} \right] \boldsymbol{\Phi} \tag{16.77}$$

In the absence of constraints, the least-squares solution of OBMPC expressed by criterion (16.74) is

$$\Delta\mathcal{U}(k|k) = [\boldsymbol{\mathcal{S}}_{H_p}^T \Gamma^T \Gamma \boldsymbol{\mathcal{S}}_{H_p} + \Lambda^T \Lambda]^{-1} \boldsymbol{\mathcal{S}}_{H_p}^T \Gamma^T \Gamma [\mathcal{R}(k+1|k) - \boldsymbol{\Phi}_{H_p} \hat{\boldsymbol{Y}}(k|k)] \tag{16.78}$$

Only $\Delta\boldsymbol{u}(k|k)$, first component of $\Delta\mathcal{U}(k|k)$, is implemented.

In Lunström et al. (1995), the residual dynamics represented by ($\boldsymbol{A}_u$, $\boldsymbol{B}_u$, $\boldsymbol{C}_u$) are approximated as $n_y \times n_u$ single-input single-output first-order systems, thus linking each input to each output. These authors also assume classically that the measurement noise v is uncorrelated white noise. Moreover, the disturbances d are assumed to be integrated white noise filtered through first-order dynamics. This allows us to obtain a very simplified Kalman filter, which avoids solving a large Riccati equation (Lee et al., 1994; Lunström et al., 1995)

16.2.8 State-Space Linear Model Predictive Control as General Optimization

A more general but less formalized optimization approach can be adopted. Consider the system described by a state-space discrete-time model in the form

$$\begin{array}{ll} \boldsymbol{x}(k+1) & = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\boldsymbol{u}(k) \\ \boldsymbol{y}(k) & = \boldsymbol{C}\boldsymbol{x}(k) \end{array} \tag{16.79}$$

and the constraints written as

$$\boldsymbol{E}\boldsymbol{x} + \boldsymbol{F}\boldsymbol{u} \leq \Psi \tag{16.80}$$

The control problem consists of solving the open-loop optimization (Morari and Lee, 1999)

$$\min_{\boldsymbol{u}} J = \boldsymbol{x}(H_p)^T S_0 \boldsymbol{x}(H_p) + \sum_{i=0}^{H_p-1} \boldsymbol{x}(i)^T Q \boldsymbol{x}(i) + \sum_{i=0}^{H_c-1} \boldsymbol{u}(i)^T R \boldsymbol{u}(i) \tag{16.81}$$

subjected to the constraints (16.80). H_p is the prediction or output horizon, H_c is the control or input horizon.

In the case where both prediction and control horizons are infinite and where no constraints are present, the linear MPC becomes the classical discrete-time LQ problem.

When the prediction horizon is finite, in the framework of MPC, this is referred to as a receding horizon control problem, as only the first control $u^*(0)$ of the optimal sequence $u^*(i)$, $i = 1, H_c - 1$, is implemented. In the case where both control horizons are finite, this is a classical optimization problem, for which available numerical subroutines exist. However, contrary to the case of the LQ control whose static state feedback control law (through the solving

of an algebraic Riccati equation) guarantees closed-loop stability, this problem is not obvious for linear MPC and several related questions are addressed by Morari and Lee (1999):

- It is possible that the constraints (16.80) render the optimization problem infeasible.
- As the optimization problem is solved in open-loop, it may occur that the closed-loop system goes out of the feasible region. In commercial algorithms (Qin and Badgwell, 1996), soft constraints exist which can be violated during some time, contrary to hard constraints. They are penalized in the objective function.
- In the case of an unstable system, in general the system cannot be stabilized globally (some states are not stabilized) when there are input saturation constraints.

This absence of stability question of the finite horizon model predictive controller was solved by using a penalty on an infinite horizon, although their number of decision variables remained finite (Muske and Rawlings, 1993; Rawlings and Muske, 1993) .

Even if the most encountered criterion to be minimized in the case of MPC is a quadratic one such as (16.31), (16.44) or (16.74), other possibilities exist. For example, Ricker (1990b) defines a linear objective function subjected to constraints, thus leading to a linear programming problem for which general-purpose packages exist. This linear objective function takes the form

$$J = \boldsymbol{w}_r^T \, |\boldsymbol{y}_r(k) - \boldsymbol{y}(k)| + \boldsymbol{w}_y^T \, \boldsymbol{y} + \boldsymbol{w}_{\Delta u}^T \, |\Delta \boldsymbol{u}(k)| + \boldsymbol{w}_u^T \, \boldsymbol{u} \tag{16.82}$$

where $\boldsymbol{w}$ are weights. In this linear objective case, stability and robustness issues can only be dealt with by simulation.

16.3 Nonlinear Model Predictive Control

16.3.1 Nonlinear Quadratic Dynamic Matrix Control (NLQDMC)

Given a nonlinear state-space model, it is tempting to linearize it around a variable operating point in order to further apply linear MPC. This is the simplest nonlinear approach, which is useful when the process to be controlled presents strong nonlinearities. It was proposed by Gattu and Zafiriou (1992, 1995) who adapted the nonlinear version of QDMC proposed by Garcia (1984) by incorporating a state estimator. Their controller can cope with unstable and integrating processes. The algorithm can be described as follows:

- Variables known at sampling instant k: measurement $y(k)$, estimation $\hat{x}(k|k-1)$ of state at time k based on information available at time $k-1$, manipulated variable $u(k-1)$, sampling period T_s.
- Objective: compute the H_c future input variations from which only the first one will be implemented.

– Step 1: Linearization of the continuous nonlinear model

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t)) \qquad , \quad \boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{u}(t)) \tag{16.83}$$

at $[\hat{\boldsymbol{x}}(k), \boldsymbol{u}(k-1)]$ to get the continuous linearized system

$$\begin{array}{l} \dot{\hat{\boldsymbol{x}}}(k|k-1) = \boldsymbol{A}_k\, \hat{\boldsymbol{x}}(k|k-1) + \boldsymbol{B}_k\, \boldsymbol{u}(k-1) \\ \boldsymbol{y}(k) = \boldsymbol{C}_k\, \boldsymbol{x}(k|k-1) + \boldsymbol{D}_k\, \boldsymbol{u}(k-1) \end{array} \tag{16.84}$$

with

$$\begin{array}{lll} \boldsymbol{A}_k = \left.\dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x},\boldsymbol{u}} & ; & \boldsymbol{B}_k = \left.\dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\right|_{\boldsymbol{x},\boldsymbol{u}} \\ \boldsymbol{C}_k = \left.\dfrac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x},\boldsymbol{u}} & ; & \boldsymbol{D}_k = \left.\dfrac{\partial \boldsymbol{h}}{\partial \boldsymbol{u}}\right|_{\boldsymbol{x},\boldsymbol{u}} \end{array} \tag{16.85}$$

– Step 2: Discretization of the linearized model (16.84) as

$$\begin{array}{l} \hat{\boldsymbol{x}}_{j+1} = \boldsymbol{\Phi}_k\, \hat{\boldsymbol{x}}_j + \boldsymbol{\Gamma}_k\, \boldsymbol{u}_j \\ \boldsymbol{y}_j = \boldsymbol{C}_k\, \boldsymbol{x}_j + \boldsymbol{D}_k\, \boldsymbol{u}_j \end{array} \tag{16.86}$$

– Step 3: Computation of the step response coefficient matrices $\boldsymbol{S}_{i,k}$ as

$$\boldsymbol{S}_{i,k} = \sum_{j=1}^{i} \boldsymbol{C}_k\, \boldsymbol{\Phi}_k^{j-1}\, \boldsymbol{\Gamma}_k \qquad (i = 1, \ldots, H_p) \tag{16.87}$$

or obtaining of these step response coefficient matrices by integration of the nonlinear system (16.83) over H_p sampling periods with initial condition $\boldsymbol{x} = 0$ and $\boldsymbol{u} = 1$. The predicted output variation due to the future input moves is

$$\Delta\hat{y}(k+l|k-1) = \sum_{i=1}^{l} \boldsymbol{S}_{i,k} \Delta\boldsymbol{u}(k+l-i) \qquad (l = 1, \ldots, H_p) \tag{16.88}$$

with $\Delta\boldsymbol{u}(k) = \boldsymbol{u}(k) - \boldsymbol{u}(k-1)$.

– Step 4: Computation of the gain $\boldsymbol{K}_k$ of the discrete Kalman linear filter (Sect. 11.1.2).

– Step 5: Prediction of the influence of past input variations. Define $\boldsymbol{y}^*(k+l|k-1)$ $(l = 1, \ldots, H_p)$ as the future output prediction assuming that all future inputs are constant and equal to $\boldsymbol{u}(k-1)$. This step is decomposed into four stages.

Stage 5-1: The predicted states $\boldsymbol{x}^*(k|k-1)$ are taken to be equal to the estimates $\boldsymbol{x}(k|k-1)$.

Stage 5-2: The disturbances at time k are $\boldsymbol{d}(k|k) = \boldsymbol{y}(k) - \boldsymbol{h}(\boldsymbol{x}(k|k-1))$ where $\boldsymbol{y}(k)$ is the measurement vector.

Stage 5-3: The future disturbances are considered to be equal to the present ones $\boldsymbol{d}(k+l|k) = \boldsymbol{d}(k|k)$ $(l = 1, \ldots, H_p)$.

Stage 5-4: Integrate sequentially the nonlinear state-space system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$ with initial condition $\boldsymbol{x}^*(k+l-1|k-1)$, $(l = 1, \ldots, H_p)$ and

$\boldsymbol{u} = \boldsymbol{u}(k-1)$ over one sampling period, then correct the result of integration by adding $\boldsymbol{K}_k\,\boldsymbol{d}$ to obtain the predicted state vector $\boldsymbol{x}^*(k+l|k-1)$. Deduce the predicted output based on the past input variations $\boldsymbol{y}^*(k+l|k-1) = \boldsymbol{h}(\boldsymbol{x}^*(k+l|k-1))$

- Step 6: Computation of the predicted output considering the influence of past and future input moves and estimated disturbances

$$\hat{\boldsymbol{y}}(k+l|k-1) = \boldsymbol{y}^*(k+l|k-1) + \sum_{i=1}^{l} \boldsymbol{S}_{i,k}\,\Delta\boldsymbol{u}(k+l-i) + \boldsymbol{d}(k|k) \qquad , \quad l = 1, \ldots, H_p \tag{16.89}$$

- Step 7: Optimization by minimization of the quadratic criterion, that is completely similar to (16.74), taking into account the predicted outputs $\hat{\boldsymbol{y}}$ on the prediction horizon from $k+1$ to $k+H_p$, while the input moves $\Delta\boldsymbol{u}$ are considered on the control horizon from k to $k+H_c$. The criterion is minimized with respect to these input variations. Gattu and Zafiriou (1992) note that this optimization problem can be formulated as a standard quadratic programming one.
- Step 8: Implementation of the first input move $\Delta\boldsymbol{u}(k)$.
- Step 9: Update of the state vector by integration over one sampling period of the nonlinear state-space system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$, with initial condition $\boldsymbol{x}^*(k|k-1)$ and just calculated $\boldsymbol{u} = \boldsymbol{u}(k)$, then correct the result of integration by adding $\boldsymbol{K}_k\,\boldsymbol{d}$ to get $\boldsymbol{x}^*(k+1|k)$.

(Gattu and Zafiriou, 1992) remark that this approach does not reject perfectly the step disturbances, but that the closed-loop system is stable in that case.

16.3.2 Other Approaches of Nonlinear Model Predictive Control

Different approaches are used in nonlinear MPC (NMPC). However, this remains a domain in development where much research is actually being carried out (Allgöwer and Zheng, 2000). The type of nonlinearity of the model can vary greatly even including second-order Volterra models (Doyle et al., 1995; Genceli and Nikolaou, 1995; Maner et al., 1996) or neural networks. In general, state-space models are required. Qin and Badgwell (2000) give an overview of NMPC applications. Interesting reviews about NMPC (Allgöwer et al., 1999; Bequette, 1991; Chen, 1997; Mayne et al., 2000; Morari and Lee, 1999) have been published, where they insist on two main obstacles concerning the extension of MPC from linear to nonlinear systems:

- the stability issue for constrained finite horizon systems. Different types of constraints may be introduced to guarantee stability for linear MPC. Similar approaches are often used for nonlinear MPC. Also an infinite or quasi-infinite horizon is used.
- the computational burden: a nonlinear optimization problem must be solved on-line and there is generally no guaranty of finding a global optimum.

Given the following nonlinear state-space model

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t)) \qquad , \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{16.90}$$

subject to input and state constraints (given in the simplest form)

$$\begin{array}{lll} \boldsymbol{x}(t) \in \mathcal{X} & , \quad \text{with:} & \mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n | \, \boldsymbol{x}_{min} \leq \boldsymbol{x} \leq \boldsymbol{x}_{max}\} \\ \boldsymbol{u}(t) \in \mathcal{U} & , \quad \text{with:} & \mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^{n_u} | \, \boldsymbol{u}_{min} \leq \boldsymbol{u} \leq \boldsymbol{u}_{max}\} \end{array} \tag{16.91}$$

for stability studies, the NMPC problem is often formulated as a finite horizon open-loop optimal control problem (Chen and Allgöwer, 1998a,b) (see Chap. 14) as

$$\min_{\hat{\boldsymbol{u}}} J(\boldsymbol{x}, \hat{\boldsymbol{u}}, T_p) = \int_t^{t+T_p} \left(\| \, \hat{\boldsymbol{x}}(\tau; \boldsymbol{x}(t), t) \, \|_Q^2 + \| \, \hat{\boldsymbol{u}}(\tau) \, \|_R^2 \right) d\tau \tag{16.92}$$

subject to

$$\dot{\hat{\boldsymbol{x}}}(t) = \boldsymbol{f}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}) \qquad , \quad \hat{\boldsymbol{x}}(t; \boldsymbol{x}, t) = \boldsymbol{x}(t) \tag{16.93}$$

and

$$\begin{array}{l} \hat{\boldsymbol{x}}(\tau; \boldsymbol{x}(t), t) \in \mathcal{X} \\ \hat{\boldsymbol{u}}(\tau) \in \mathcal{U} \end{array} \tag{16.94}$$

$\|\boldsymbol{x}\|_Q^2 = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}$ is the Euclidean norm weighted by the positive definite matrix $\boldsymbol{Q}$ (same for $\boldsymbol{R}$). T_p is the prediction horizon. The control horizon T_c not mentioned here is such that $T_c \leq T_p$. $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{x}}$ are the predicted variables by the controller differing from the actual values $\boldsymbol{u}$ and $\boldsymbol{x}$ on the plant.

For normal use, the functional F under the integral would be formulated as

$$F(\boldsymbol{x}, \boldsymbol{u}) = (\boldsymbol{x} - \boldsymbol{x}^r)^T \, \boldsymbol{Q} \, (\boldsymbol{x} - \boldsymbol{x}^r) + (\boldsymbol{u} - \boldsymbol{u}^r)^T \, \boldsymbol{R} \, (\boldsymbol{u} - \boldsymbol{u}^r) \tag{16.95}$$

where $\boldsymbol{x}^r$ and $\boldsymbol{u}^r$ are reference trajectories. However, for the stability study, it suffices to consider $(\boldsymbol{x}^r, \boldsymbol{u}^r) = (0, 0)$ as the steady-state point.

Because of the finite horizon in (16.92), closed-loop stability cannot be guaranteed in general, although it can be obtained by appropriate parameter tuning of an MPC scheme. It might be possible to use an infinite horizon as in LQG control to guarantee closed-loop stability; however, this would result in an infinite or high-dimension optimization problem, which is not desirable. For that reason, researchers have suggested different solutions to solve the closed-loop stability problem.

Mayne and Michalska (1990) introduced a terminal equality constraint

$$\hat{\boldsymbol{x}}(t + T_p) = 0 \tag{16.96}$$

which forces the state to zero at the end of the prediction horizon. The optimal objective function can then be considered as a Lyapunov function. The same authors (Michalska and Mayne, 1993) relaxed this condition by transforming it into a terminal inequality constraint

$$\hat{\boldsymbol{x}}(t + T_p) \in \Omega \tag{16.97}$$

where Ω is a terminal region which is a region of attraction for the nonlinear system controlled locally by a linear state feedback law: $\boldsymbol{u} = \boldsymbol{K}\,\boldsymbol{x}$. Thus, a dual-mode controller is proposed as a receding horizon controller: outside the terminal region, a receding horizon controller is applied and inside the terminal region, the linear state feedback law, so that switching between the two controllers must be performed. To be feasible, the state at the end of the finite horizon must be on the boundary of the terminal region. The following optimization problem must then be solved

$$\min_{\hat{\boldsymbol{u}}, T_p} J(\boldsymbol{x}, \hat{\boldsymbol{u}}, T_p) \tag{16.98}$$

where T_p (equal to T_c) becomes a minimization variable.

Chen and Allgöwer (1998a,c) added a terminal penalty term similarly to Michalska and Mayne (1993) to the finite horizon objective function which becomes

$$\min_{\hat{\boldsymbol{u}}} J(\boldsymbol{x}, \hat{\boldsymbol{u}}, T_p) = \|\hat{\boldsymbol{x}}(t+T_p; \boldsymbol{x}(t), t)\|_P^2 + \int_t^{t+T_p} \left(\|\hat{\boldsymbol{x}}(\tau; \boldsymbol{x}(t), t)\|_Q^2 + \|\hat{\boldsymbol{u}}(\tau)\|_R^2 \right) d\tau \tag{16.99}$$

to obtain the quasi-infinite NMPC which approximates to infinite horizon prediction. The constraints are now

$$\begin{array}{l} \dot{\hat{\boldsymbol{x}}}(t) = \boldsymbol{f}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}) \quad , \quad \hat{\boldsymbol{x}}(t; \boldsymbol{x}, t) = \boldsymbol{x}(t) \\ \hat{\boldsymbol{x}}(\tau; \boldsymbol{x}(t), t) \in \mathcal{X} \quad , \quad \tau \in [t, t+T_p] \\ \hat{\boldsymbol{u}}(\tau) \in \mathcal{U} \quad , \quad \tau \in [t, t+T_p] \\ \hat{\boldsymbol{x}}(t+T_p; \boldsymbol{x}(t), t) \in \Omega \end{array} \tag{16.100}$$

When the quasi-infinite horizon objective function is written in more general terms than the quadratic terms used in (16.99), finding the terminal region is, in general, very difficult. In the case of the quadratic expression (16.99) and when the system admits a linearizable part that is controllable, Chen and Allgöwer (1998b) proposed the following off-line procedure (very similar to Michalska and Mayne (1993)) in order to compute the terminal region:

Step 1: Calculate the Jacobian linearization $(\boldsymbol{A}, \boldsymbol{B})$ of (16.90), then determine the locally stabilizing linear state feedback: $\boldsymbol{u} = \boldsymbol{K}\,\boldsymbol{x}$

Step 2: choose a positive constant $\alpha < -\lambda_{max}(\boldsymbol{A}_K)$ and solve the Lyapunov equation

$$(\boldsymbol{A}_K + \alpha\,\boldsymbol{I})^T\,\boldsymbol{P} + \boldsymbol{P}\,(\boldsymbol{A}_K + \alpha\,\boldsymbol{I}) = -(\boldsymbol{Q} + \boldsymbol{K}^T\,\boldsymbol{R}\,\boldsymbol{K}) \tag{16.101}$$

to obtain the positive definite matrix $\boldsymbol{P}$, with $\boldsymbol{A}_K = \boldsymbol{A} + \boldsymbol{B}\,\boldsymbol{K}$.

Step 3: Find the largest β_1 defining the region Ω_1 such that the state and input constraints are satisfied for $\boldsymbol{x} \in \Omega_1$

$$\begin{array}{l} \Omega_1 = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x} \leq \beta_1\} \\ \text{and} \quad \Omega_1 \subseteq \mathcal{X} \qquad \text{and} \quad \boldsymbol{K}\boldsymbol{x} \in \mathcal{U} \qquad \forall \boldsymbol{x} \in \Omega_1 \end{array} \tag{16.102}$$

Step 4: Find the largest $\beta \in]0, \beta_1]$ specifying a terminal region Ω

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^n \,|\, \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x} \leq \beta\} \tag{16.103}$$

such that the optimal state solution of the following optimization problem is non-positive

$$\max_{\boldsymbol{x}} \left\{ \boldsymbol{x}^T \boldsymbol{P} \phi(\boldsymbol{x}) - \alpha \, \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x} \,|\, \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x} \leq \beta \right\} \tag{16.104}$$

with: $\phi(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{K}\boldsymbol{x}) - \boldsymbol{A}_K \boldsymbol{x}$.

Step 5: Choose the prediction horizon T_p satisfying

$$T_p \geq T_c + T_s \tag{16.105}$$

where T_s is maximum time needed for the uncontrolled system to reach Ω starting from $\boldsymbol{x}_0$. At most, T_s could be chosen as the settling time.

The approaches that have been previously described share the objective of satisfying closed-loop stability and possible real-time implementation. Other approaches closer to optimization are possible especially, for large-scale systems (Biegler, 2000; Doyle et al., 1997; Van Antwerp and Braatz, 2000). For example, a direct multiple shooting method that is applied on-line to a large scale system, i.e. a high-purity distillation column has been proposed (Bock et al., 2000; Diehl et al., 2001).

16.4 Model Predictive Control of a FCC

16.4.1 FCC Modelling

Description of the Fluid Catalytic Cracking Process

The fluid catalytic cracking (FCC) unit contains three main parts: the reactor riser, the particle separator and the catalyst regenerator. The catalytic cracking of petroleum fractions provides a large range of products including gasoline, light olefins, gas oil, paraffins, naphthenes, aromatics and olefinic polymers. The FCC process has been used with many evolutions for more than half a century (Avidan and Shinnar, 1990).

The riser is considered as a plug flow reactor. The space time for the riser is a few seconds. Due to coking in the riser, the catalyst loses its activity in a few seconds. The cracking reaction is endothermic.

After the riser, a larger vessel, called the separator, is situated which has two roles: this vessel hosts the riser cyclones and stripping steam enters this particle separator so that the catalyst is separated from the gaseous products. The separator is considered as a CSTR.

The regenerator is designed to regenerate the catalyst by combustion of the coke. It is a fluidized bed composed of two zones: the dense zone, where most of the catalyst is present, and the dilute zone, where practically no catalyst is present. Some authors consider that the dense zone is divided into two phases:

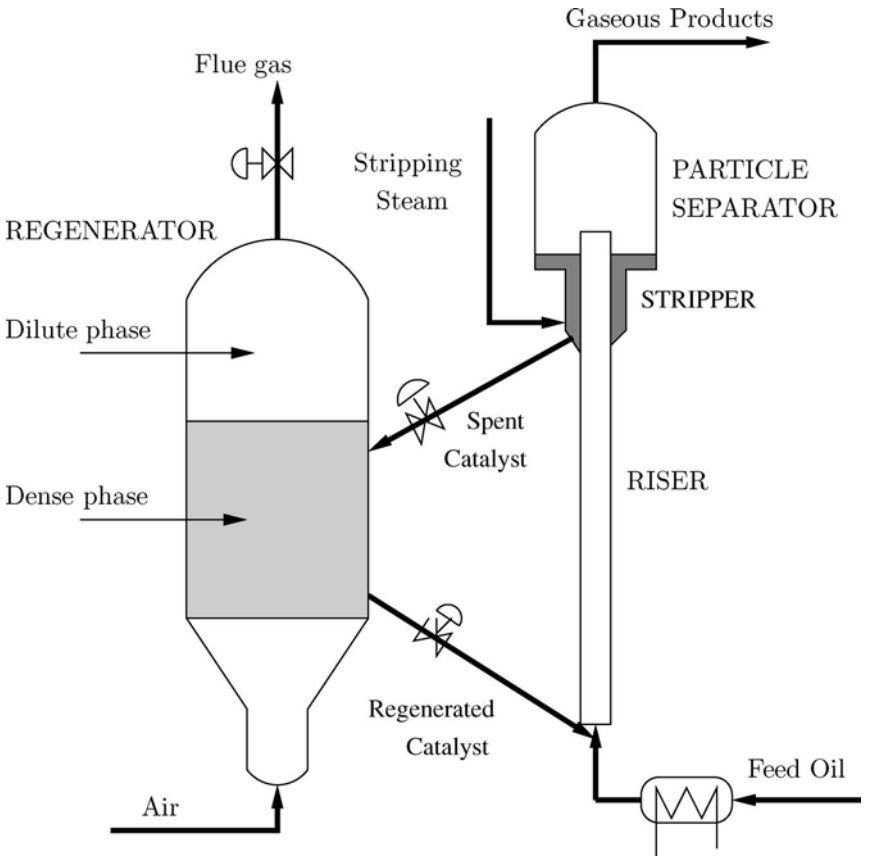


Figure 16.3: Scheme of a fluid catalytic cracking unit

the emulsion phase, which acts as a CSTR, and the bubble phase, which acts as a plug flow. A good description of the behaviour of the regenerator considered as a fluidized bed can be found in literature (Kunii and Levenspiel, 1991).

The importance of FCC control is emphasized by many authors (Arbel et al., 1996; Grosdidier et al., 1993). The process dynamics are complex. The FCC optimization is primordial for the refinery economics. The choice of the control structure is important. Moreover, operating constraints and the nonlinear behaviour of the process make the process control problem very attractive for performing multivariable algorithms such as MPC ones.

A Short Review of FCC Models

Most of the FCC literature studies are based on empirical or semi-empirical models, which usually conform quite well with the real industrial process in a small operating range. However, when the operating conditions change, their validity can fail. So these types of models are unreliable. Many models do

not describe important variable variations such as pressure effects (Balchen et al., 1992; Hovd and Skogestad, 1993), use oversimplified kinetics or very approximate models (Ansari and Tadé, 1997; Christofides and Daoutidis, 1997; Khandalekar and Riggs, 1995). In most published papers, the greatest differences deal with the modelling of the dense bed in the regenerator. Although the bubbling bed model (Kunii and Levenspiel, 1991) is often referred to, due to the complexities of the proposed approaches and some lack of information, the degree of refinement of the regenerator model very much differs according to the authors (Arbel et al., 1995a). Authors even disagree on the necessity of taking into account the spatial character for the bubble phase in the dense bed (Lasa et al., 1981) although this is more realistic and considered by more detailed papers (McFarlane et al., 1993). The freeboard over the dense bed is not often considered in the modelling. Of course, these differences concerning the FCC models have strong implications on the frequently related instability issues (Arandes and de Lasa, 1992; Arbel et al., 1995b; Elnashaie and Elshishini, 1993; Elshishini and Elnashaie, 1990). On the other hand, a full model such as the one proposed by Han and Chung (2001a,b) probably incorporates too much complexity for the objective of control itself. In particular, partial differential equations are not always necessary when time scales have largely different orders. Thus, the question of an adequately reduced model is crucial and gives rise to different models with regard to the pursued objective.

A good compromise is the reliable model of moderate complexity which has been proposed by Rohani and co-workers (Ali et al., 1997; Ali, 1995; Malay, 1998). In their studies, a four-lump reaction scheme (Lee et al., 1989) is used for the riser reactions, the bubbling bed Kunii-Levenspiel for the regenerator with ordinary differential equations for the dense phase and spatial effects for the gaseous species in the bubble phase; the pressure effects have been incorporated.

Description of the Considered FCC Model

The model is adapted from those used by Balchen et al. (1992) and Hovd and Skogestad (1993). In their study, the spent and regenerated catalyst flow rates are equal although they are not in real FCCs. Pressure effects are not taken into account. Their riser model is inspired from Lee and Groves (1985) who used a three-lump kinetics. Their regenerator model comes from Errazu et al. (1979). In the present model, only the case of partial combustion of coke in the regenerator is considered. Typical constraints for FCC variables are commented in the literature Monge and Georgakis (1987).

Riser:

Balchen et al. (1992) avoided the use of spatial derivatives by introducing intermediate variables at a specified height in the riser. We have checked that this allows to obtain very similar results to the classical case where spatial derivatives are used. However, the latter usage is more immediate and more common so it was preferred.

The feed temperature $T_{ris}(z = 0)$ at the riser inlet is derived from an energy

balance at the inlet

$$F_{cat}\,C_{p,cat}\,(T_{reg}-T_{ris}(0)) = \begin{array}{l} F_{feed}\,[C_{p,ol}\,(T_{boil}-T_{feed})+ \\ \Delta H_{vap}+C_{p,og}\,(T_{ris}(0)-T_{boil})] \end{array} \tag{16.106}$$

resulting in

$$T_{ris}(0) = \frac{F_{cat}\,C_{p,cat}\,T_{reg} - F_{feed}\,(C_{p,ol}\,(T_{boil}-T_{feed}) + \Delta H_{vap} - C_{p,og}\,T_{boil})}{F_{cat}\,C_{p,cat} + F_{feed}\,C_{p,og}} \tag{16.107}$$

In the riser, only two components are considered: gas oil and gasoline. For a more complex lumped model, similar equations would be obtained. The kinetic constants follow Arrhenius law, for gas oil consumption

$$k_1 = k_{10}\,\exp(-E_{af}/(R\,T_{ris})) \tag{16.108}$$

and for gasoline production

$$k_3 = k_{30}\,\exp(-E_{ag}/(R\,T_{ris})) \tag{16.109}$$

The catalyst undergoes a deactivation along the riser

$$\phi = (1-m\,C_{coke,reg})\,\exp(-\alpha\,t_c\,z\,C_{owr}) \qquad \text{with:} \quad C_{owr} = F_{cat}/F_{feed} \tag{16.110}$$

Two mass balances give the following respective spatial equations for gas oil and gasoline weight fractions

$$\begin{array}{l} \dfrac{dy_{go}}{dz} = -k_1\,y_{go}^2\,C_{owr}\,\phi\,t_c \\ \dfrac{dy_g}{dz} = (\alpha_2\,k_1\,y_{go}^2 - k_3\,y_g)\,C_{owr}\,\phi\,t_c \end{array} \tag{16.111}$$

while the energy balance gives the temperature variation along the riser

$$\frac{dT_{ris}}{dz} = \frac{\Delta H_{r,crack}\,F_{feed}}{F_{cat}\,C_{p,cat} + F_{feed}\,C_{p,ol} + \lambda\,F_{feed}\,C_{p,steam}}\,\frac{dy_{go}}{dz} \tag{16.112}$$

The temperature at the exit of the riser is denoted by $T_{ris}(1)$. The concentration of coke produced is

$$C_{coke,prod} = k_c\,\sqrt{t_c\,\exp\left(-\frac{E_{acf}}{R\,T_{ris}(1)}\right)} \tag{16.113}$$

so that the concentration of coke leaving the riser is

$$C_{coke,ris}(1) = C_{coke,reg} + C_{coke,prod} \tag{16.114}$$

Separator

A separator follows the riser simply with a mass balance for the coke on the catalyst

$$\frac{dC_{coke,sep}}{dt} = \frac{F_{cat}\,(C_{coke,ris}(1) - C_{coke,sep})}{m_{cat,sep}} \tag{16.115}$$

and an energy balance giving the temperature variation

$$\frac{dT_{sep}}{dt} = \frac{F_{cat}\, C_{p,cat}\, (T_{ris}(1) - T_{sep})}{m_{cat,sep}\, C_{p,cat}} \tag{16.116}$$

Note that $T_{ris}(1) = T_{sep}$ and $C_{C,ris}(1) = C_{coke,sep}$.

Regenerator

Complex models describing the regenerator behaviour in detail are available in the literature. Here, a very simplified model is used. The model assumes partial combustion of coke, i.e. remains of CO leave the regenerator.

First, the molar ratio σ of CO_2 to CO in the dense bed is introduced as an empirical relation

$$\begin{array}{lll} \sigma = 0.000953\, \exp(5585/T_{reg}) & \text{if} & T_{reg} < 803 \\ \sigma = 1 + (T_{reg} - 803)\, 0.00142 & \text{if} & 803 < T_{reg} < 873 \\ \sigma = 1.1 + (T_{reg} - 873)\, 0.0061 & \text{if} & 873 < T_{reg} \end{array} \tag{16.117}$$

The heat of combustion of coke is equal to

$$\Delta H_{cb} = -\Delta H_1 - \Delta H_2\, (T_{reg} - 960) + 0.6\, (T_{reg} - 960)^2 \tag{16.118}$$

The rate of coke combustion (kg/s) is equal to

$$r_{cb} = k_{cb}\, \exp(-E_{acb}/(R\, T_{reg}))\, x_{O_2}\, C_{coke,reg}\, m_{cat,reg} \tag{16.119}$$

A first mass balance gives the derivative of coke on the catalyst

$$\frac{dC_{coke,reg}}{dt} = \frac{F_{cat}\, (C_{coke,sep} - C_{coke,reg}) - r_{cb}}{m_{cat,reg}} \tag{16.120}$$

and a second mass balance gives the derivative of mole fraction of O_2 in the dense bed

$$\frac{dx_{O_2}}{dt} = \frac{1}{m_{air,reg}} \left[\frac{F_{air}}{M_{w,air}}\, (x_{O_2,in} - x_{O_2}) - \frac{(1+\sigma)\, n_{CH} + 2 + 4\,\sigma}{4\,(1+\sigma)}\, \frac{r_{cb}}{M_{w,coke}} \right] \tag{16.121}$$

The energy balance gives the temperature variation

$$\begin{aligned} \frac{dT_{reg}}{dt} = \frac{1}{m_{cat,reg}\, C_{p,cat}} \big[& F_{cat}\, C_{p,cat}\, T_{sep} + F_{air}\, C_{p,air}\, T_{air} \\ & - (F_{cat}\, C_{p,cat} + F_{air}\, C_{p,air})\, T_{reg} - \Delta H_{cb}\, \frac{r_{cb}}{M_{w,coke}} \big] \end{aligned} \tag{16.122}$$

Thus, the FCC model is constituted by seven time-ordinary differential equations. The spatial equations describing the riser are integrated at each time step and the corresponding variables are then updated.

The values of the different parameters necessary for model simulation are given in Table 16.1 and the steady-state values are given in Table 16.2.

The manipulated inputs are the flow rate of regenerated catalyst u_1 and the flow rate of air u_2. The controlled outputs are the temperature at the outlet of

Table 16.1: Parameters for FCC model

Symbol	Significance	Value
F_{cat}	Mass flow rate of catalyst (kg.s^{-1})	294
F_{feed}	Mass flow rate of oil in the feed (kg.s^{-1})	40.63
F_{air}	Mass flow rate of air to regenerator (kg.s^{-1})	25.378
T_{air}	Temperature of air to regenerator (K)	360
T_{feed}	Temperature of feed (K)	434.63
T_{boil}	Boiling temperature of feed (K)	700
ΔH_{vap}	Heat of vaporization of oil (J.kg^{-1})	1.56×10^5
$C_{p,ol}$	Heat capacity of oil as a liquid (J.kg^{-1}.K^{-1})	2671
$C_{p,og}$	Heat capacity of oil as a gas (J.kg^{-1}.K^{-1})	3299
E_{acf}	Activation energy for coke formation (J.mol^{-1})	41.79×10^3
n_{prod}	Exponent in expression of coke produced	0.4
t	Time (s)	
t_c	Catalyst time residence in riser (s)	9.6
z	Dimensionless height in the riser	
k_c	Rate constant for catalytic coke formation (s$^{-0.5}$)	0.0176
$C_{p,air}$	Heat capacity of air (J.kg^{-1}.K^{-1})	1074
$C_{p,cat}$	Heat capacity of catalyst (J.kg^{-1}.K^{-1})	1005
$C_{p,steam}$	Heat capacity of dispersing steam (J.kg^{-1}.K^{-1})	1900
λ	Weight fraction of steam in feed stream to riser	0.035
k_{10}	Reaction rate constant for cracking of gas oil	9.65×10^5
k_{30}	Reaction rate constant for cracking of gasoline	4.22×10^5
E_{af}	Activation energy for gas oil cracking (J.mol^{-1})	101.5×10^3
E_{ag}	Activation energy for gasoline cracking (J.mol^{-1})	112.6×10^3
m	Empirical deactivation parameter	80
α	Catalyst decay rate constant (s^{-1})	0.12
α_2	Fraction of the gas oil that cracks to gasoline	0.75
$\Delta H_{r,crack}$	Heat of reaction for gas oil cracking (J.kg^{-1})	506.2×10^3
k_{cb}	Reaction rate constant for coke combustion	2.077×10^8
E_{acb}	Activation energy for coke combustion (J.mol^{-1})	158.59×10^3
$m_{cat,reg}$	Holdup of solid in regenerator (kg)	175738
$m_{cat,sep}$	Holdup of solid in separator (kg)	17500
$m_{air,reg}$	Holdup of air in regenerator (mol)	20000
n_{CH}	Number of hydrogen in coke of formula CH_n	2
$M_{w,coke}$	Molecular weight of coke (kg.mol^{-1})	14×10^{-3}
ΔH_1	Parameter in heat of reaction for coke combustion	521.15×10^3
ΔH_2	Parameter in heat of reaction for coke combustion	245

riser $T_{ris}(1)$ and the temperature in the regenerator T_{reg}. The sampling time is taken to be equal to 250 s.

To find the steady states, first a nonlinear optimization is realized to obtain

Table 16.2: Steady-state values of main variables

Symbol	Significance	Value
$T_{ris}(0)$	Temperature in the riser at $z = 0$ (K)	805.48
$T_{ris}(1)$	Temperature in the riser at $z = 1$ (K)	780.68
T_{reg}	Temperature in the dense bed of the regenerator (K)	972
$C_{C,reg}$	Coke concentration in the regenerator (kg/kg)	0.0038
$C_{C,ris}(1)$	Coke concentration in the riser at $z = 1$ (kg/kg)	0.01045
x_{O_2}	Oxygen mole fraction in the regenerator	0.0047
y_{go}	Weight fraction of gas oil in the riser at $z = 1$	0.4825
y_g	Weight fraction of gasoline in the riser at $z = 1$	0.3680

the values of the flow rate of air F_{air}, the feed temperature T_{feed} and the rate constant for catalytic coke formation k_c so that the regenerator steady-state balances are verified. k_c depends on the feed oil composition. Then, a dynamic simulation based on these values is realized, which confirms the steady state.

Open-Loop Responses

Open-loop responses have been obtained for step variations of 5% of the manipulated inputs around the steady state (Figs. 16.4 and 16.5).

For steps of the air flow rate, the responses are classical. However, for steps of the catalyst flow rate, the temperature at the riser exit first shows a sharp transition because of the direct influence of the concerned manipulated input on the studied output, then an inverse response. The regenerator temperature also shows an inverse response.

Quadratic Dynamic Model Control of the FCC

The model horizon is $H_m = 60$, the prediction horizon $H_p = 20$ and the control horizon $H_c = 3$. To avoid the influence of the inverse responses, a minimum bound is imposed on the prediction equal to $H_{min} = 10$ so that the predicted outputs considered for optimization are between H_{min} and H_p. The inputs and outputs considered in the quadratic criterion are normalized so that simple weights Γ and Λ can be considered.

First case: the weights on the inputs and outputs are all equal to 1. Although this is the simplest set of weights which can be used, the outputs satisfactorily follow the step set points and the inputs vary smoothly (Fig. 16.6).

Second case: the weights on the inputs are all equal to 1 and on the outputs all equal to 10. The tracking is improved with respect to the first case at the expense of larger, although quite acceptable, input variations (Fig. 16.7).

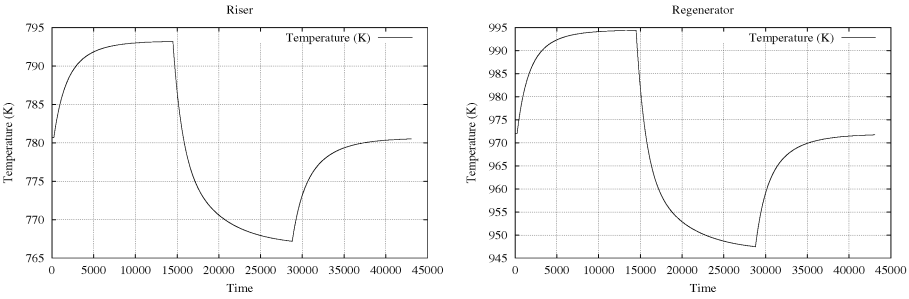


Figure 16.4: Variations of the temperature at the outlet of riser $T_{ris}(1)$ and the temperature in the regenerator T_{reg} for air flow rate steps of 5%

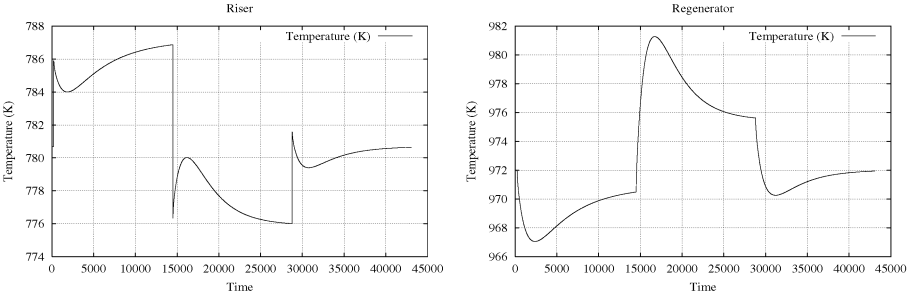


Figure 16.5: Variations of the temperature at the outlet of riser $T_{ris}(1)$ and the temperature in the regenerator T_{reg} for catalyst flow rate steps of 5%

Observer-based Model Predictive Control of the FCC

The observer-based model predictive control (OBMPC) described in Sect. 16.2.7 has been applied and its performances can be compared to the QDMC strategy. An important difference is that in OBMPC, a linear Kalman filter used as a state observer is implemented, thus simulated measurements are performed. It has been assumed that the temperatures at the riser exit and in the regenerator are random variables following the normal law with a given standard deviation. The noise is blank, i.e. its mean is zero. The model horizon is $H_m = 60$, the prediction horizon $H_p = 20$ and the control horizon $H_c = 3$. The covariance matrix $\boldsymbol{Q}$ of the Kalman filter is scalar and equal to the identity matrix. The covariance matrix $\boldsymbol{R}$ of the Kalman filter is also scalar, with a factor equal to the variance of the measurements. The initial states have been taken to be equal to the steady-state outputs. Normalized weights Γ and Λ have been taken in the quadratic criterion.

First case: to be able to make a comparison with QDMC, the standard deviation of the measurements has been taken to be equal to 0.01 so that the measurement noise is negligible. The resulting outputs obtained with OBMPC (Fig. 16.8) are very similar to those obtained with QDMC (Fig. 16.6) in similar conditions except for the presence of measurements. The weights on

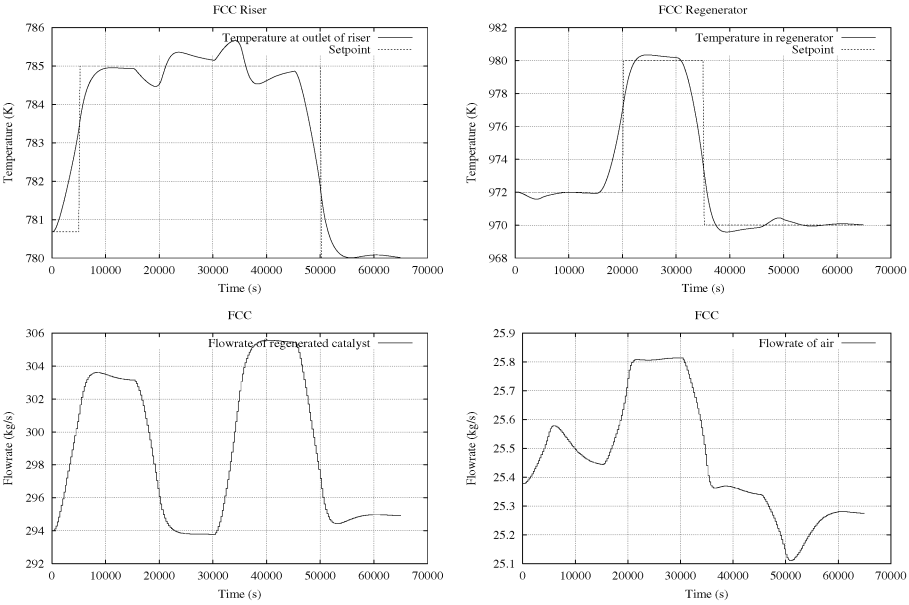


Figure 16.6: First case of quadratic dynamic matrix control of the FCC. Top: controlled outputs. Bottom: manipulated inputs

the outputs were equal to 10 while they were 1 on the inputs.

Second case: the standard deviation of the measurements has been taken to be equal to 0.5 so that the measurement noise is noticeable. The normalized weights on the outputs and inputs were taken to be equal to 1, in order to avoid brutal input moves. The resulting measured outputs obtained with OBMPC (Fig. 16.9) correctly follow the step set points in spite of the present noise. The corresponding inputs are acceptable and do not show sharp variations, in general, as expected from the chosen weights.

References

A. Al-Ghazzawi, E. Ali, A. Nouh, and E. Zafiriou. On-line tuning strategy for model predictive controllers. *J. Proc. Cont.*, 11:265–284, 2001.

H. Ali, S. Rohani, and J.P. Corriou. Modelling and control of a riser type fluid catalytic cracking (FCC) unit. *Trans. IChemE.*, 75, part A:401–412, 1997.

H.K.A. Ali. *Dynamic Modeling and Control of a Riser-Type Fluid Catalytic Cracking Unit.* Phd thesis, University of Saskatchewan, Saskatoon, Canada, 1995.

F. Allgöwer and A. Zheng, editors. *Nonlinear Model Predictive Control.* Birkhäuser, Basel, 2000.

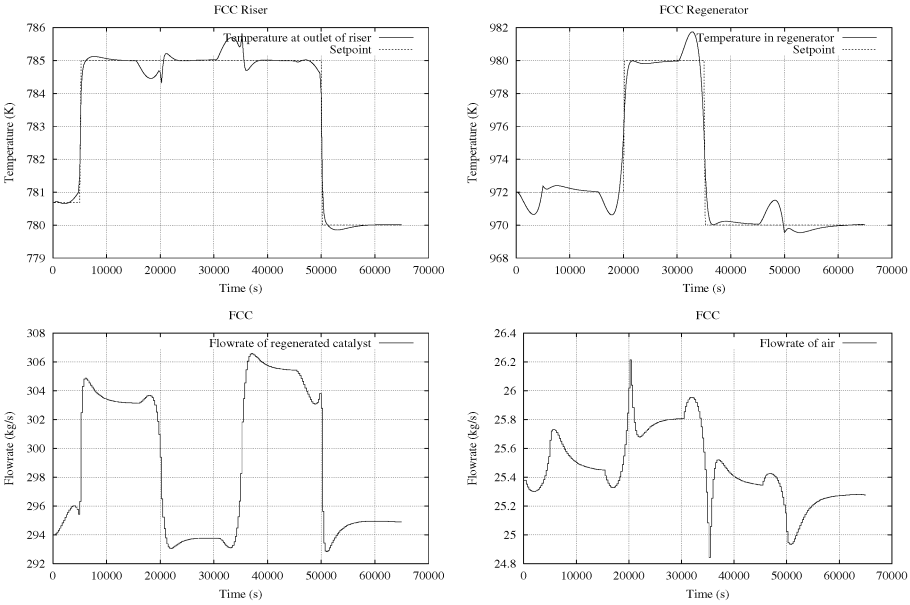


Figure 16.7: Second case of quadratic dynamic matrix control of the FCC. Top: controlled outputs. Bottom: manipulated inputs

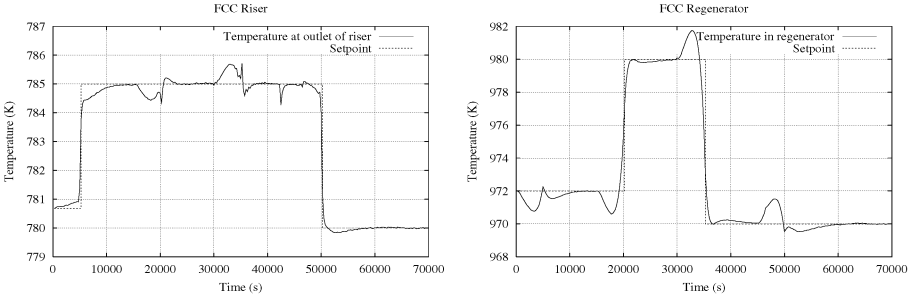


Figure 16.8: First case of observer-based model predictive control of the FCC. Variations of the controlled outputs

F. Allgöwer, T.A. Badgwell, J.S. Qin, J.B. Rawlings, and S.J. Wright. Nonlinear predictive control and moving horizon estimation - An introductory overview. In P. M. Frank, editor, *Advances in Control*, pages 391–449. Springer-Verlag, Berlin, 1999.

R.M. Ansari and M.O. Tadé. Constrained nonlinear multivariable control of a fluid catalytic cracking process. *J. Proc. Cont.*, 10:539–555, 1997.

J.M. Arandes and H.I. de Lasa. Simulation and multiplicity of steady states in fluidized FCCUs. *Chem. Eng. Sci.*, 47(9–11):2535–2540, 1992.

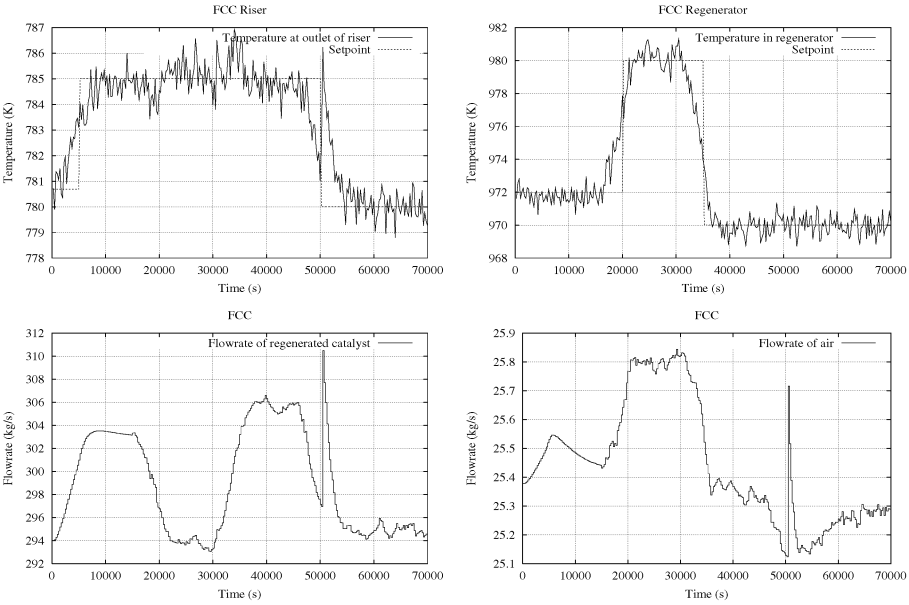


Figure 16.9: Second case of observer-based model predictive control of the FCC. Top: controlled outputs. Bottom: manipulated inputs

A. Arbel, Z. Huang, I.H. Rinard, R. Shinnar, and A.V. Sapre. Dynamic and control of fluidized catalytic crackers. 1. Modelling of the current generation of FCC's. *Ind. Eng. Chem. Res.*, 34:1228–1243, 1995a.

A. Arbel, Z. Huang, I.H. Rinard, R. Shinnar, and A.V. Sapre. Dynamic and control of fluidized catalytic crackers. 2. Multiple steady states and instabilities. *Ind. Eng. Chem. Res.*, 34:3014–3026, 1995b.

A. Arbel, I.H. Rinard, and R. Shinnar. Dynamic and control of fluidized catalytic crackers. 3. Designing the control system: choice of manipulated and measured variables for partial control. *Ind. Eng. Chem. Res.*, 35:2215–2233, 1996.

A.A. Avidan and R. Shinnar. Development of catalytic cracking technology. a lesson in chemical reactor design. *Ind. Eng. Chem. Res.*, 29:931–942, 1990.

J.S. Balchen, D. Ljungquist, and S. Strand. State-space predictive control. *Chem. Eng. Sci.*, 47(4):787–807, 1992.

B.W. Bequette. Nonlinear control of chemical processes: a review. *Ind. Eng. Chem. Res.*, 30:1391–1413, 1991.

L. Biegler. Efficient solution of dynamic optimization and NMPC problems. In F. Allgöwer and A. Zheng, editors, *Nonlinear Predictive Control*. Birkhäuser, 2000.

R. R. Bitmead, M. Gevers, and V. Wertz. *Adaptive Optimal Control, The Thinking Man's GPC*. Prentice Hall, New York, 1990.

H. G. Bock, M. Diehl, J.P. Schlöder, F. Allgöwer, R. Findeisen, and Z. Nagy. Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. In *International Symposium on Advanced Control of Chemical Processes*, pages 695–702, Pisa Italy, 2000.

E.F. Camacho and C. Bordons. *Model Predictive Control in the Process Industry*. Springer-Verlag, Berlin, 1995.

E.F. Camacho and C. Bordons. *Model Predictive Control*. Springer-Verlag, Berlin, 1998.

H. Chen. *Stability and Robustness Considerations in Nonlinear Model Predictive Control*. PhD thesis, Stuttgart University, 1997.

H. Chen and F. Allgöwer. A computationally attractive nonlinear predictive control scheme with guaranteed stability for stable systems. *J. Proc. Cont.*, 8(5-6):475–485, 1998a.

H. Chen and F. Allgöwer. Nonlinear model predictive control schemes with guaranteed stability. In R. Berber and C. Kravaris, editors, *Nonlinear Model Based Process Control*, pages 465–494. Kluwer Academic Press, Netherlands, 1998b.

H. Chen and F. Allgöwer. A quasi-infinite horizon nonlinear model predictive control scheme with guaranteed stability. *Automatica*, 34(10):1205–1218, 1998c.

P.D. Christofides and P. Daoutidis. Robust control of multivariable two-time scale nonlinear systems. *J. Proc. Cont.*, 7(5):313–328, 1997.

D.W. Clarke. *Advances in Model-Based Predictive Control*. Oxford University Press, Oxford, 1994.

D.W. Clarke, C. Mohtadi, and P.S. Tuffs. Generalized predictive control - Part I. The basic algorithm. *Automatica*, 23(2):137–148, 1987a.

D.W. Clarke, C. Mohtadi, and P.S. Tuffs. Generalized predictive control - Part II. Extensions and interpretations. *Automatica*, 23(2):149–160, 1987b.

C.R. Cutler and B.L. Ramaker. Dynamic matrix control - a computer control algorithm. In *AIChE Annual Meeting*, Houston, Texas, 1979.

H. Demircioglu and P.J. Gawthrop. Continuous-time generalized predictive control (CGPC). *Automatica*, 27(1):55–74, 1991.

H. Demircioglu and P.J. Gawthrop. Multivariable continuous-time generalized predictive control (MCGPC). *Automatica*, 28(4):697–713, 1992.

M. Diehl, I. Uslu, R. Findeisen, S. Schwartzkopf, F. Allgöwer, H.G. Bock andT. Bürner, E.D. Gilles, A. Kienle, J.P. Schlöder, and E. Stein. Real-time optimization for large scale processes: nonlinear model predictive control of a high-purity distillation column. In Groeschel, Krumke, and Rambau, editors, *Online Optimization of Large Scale Systems: State of the Art*, pages 363–383. Springer, 2001.

F.J. Doyle, B.A. Ogunnaike, and R.K. Pearson. Nonlinear model based control using second order Volterra models. *Automatica*, 31(5):697–714, 1995.

F.J. Doyle, J.F. Pekny, P. Dave, and S. Bose. Specialized programming methods in the model predictive control of large-scale systems. *Comp. Chem. Engng.*, 21:s847–852, 1997.

S.S.E.H. Elnashaie and S.S. Elshishini. Digital simulation of industrial fluid catalytic cracking units-IV dynamic behaviour. *Chem. Eng. Sci.*, 1993.

S.S. Elshishini and S.S.E.H. Elnashaie. Digital simulation of industrial fluid catalytic cracking units: bifurcation and its implications. *Chem. Eng. Sci.*, 1990.

A.F. Errazu, H.I. de Lasa, and F. Sarti. A fluidized bed catalytic cracking regenerator model grid effects. *Can. J. Chem. Engng.*, 57:191–197, 1979.

R. Fletcher. *Practical Methods of Optimization.* Wiley, Chichester, 1991.

J.B. Froisy. Model predictive control: past, present and future. *ISA Transactions*, 33:235–243, 1994.

C.E. Garcia. Quadratic dynamic matrix control of nonlinear processes - an application to a batch reaction process. In *AIChE Annual Meeting*, San Franciso, USA, 1984.

C.E. Garcia and M. Morari. Internal model control. 1. A unifying review and some new results. *Ind. Eng. Chem. Process Des. Dev.*, 21:308–323, 1982.

C.E. Garcia and A.M. Morshedi. Quadratic programming solution of dynamic matrix control (QDMC). *Chem. Eng. Comm.*, 46:73–87, 1986.

C.E. Garcia, D.M. Prett, and M. Morari. Model predictive control: Theory and practice - a survey. *Automatica*, 25(3):335–348, 1989.

G. Gattu and E. Zafiriou. Nonlinear quadratic dynamic matrix control with state estimation. *Ind. Eng. Chem. Res.*, 31:1096–1104, 1992.

G. Gattu and E. Zafiriou. Observer based nonlinear quadratic dynamic matrix control for state space and input/output models. *Can. J. Chem. Eng.*, 73: 883–895, 1995.

H. Genceli and M. Nikolaou. Robust stability analysis of constrained l_1-norm model predictive control. *AIChE J.*, 39(12):1954–1965, 1993.

H. Genceli and M. Nikolaou. Design of robust constrained model-predictive controllers with Volterra series. *AIChE J.*, 41(9):2098–2107, 1995.

P. Grosdidier, A. Mason, A. Aitolahti, P. Heinonen, and V. Vanhamäki. FCC unit reactor-regenerator control. *Comp. Chem. Engng.*, 17(2):165–179, 1993.

P.H. Gusciora, J.H. McAmis, R.C. Sorensen, and C.R. Cutler. Experiences applying DMC on a model IV FCC. In *paper 131L, AIChE meeting*, Miami, Fl, 1992.

I.S. Han and C.B. Chung. Dynamic modeling and simulation of a fluidized catalytic cracking process. Part I: Process modeling. *Chem. Eng. Sci.*, 56: 1951–1971, 2001a.

I.S. Han and C.B. Chung. Dynamic modeling and simulation of a fluidized catalytic cracking process. Part II: Property estimation and simulation. *Chem. Eng. Sci.*, 56:1973–1990, 2001b.

M. Hovd and S. Skogestad. Procedure for regulatory control structure selection with application to the FCC process. *AIChE J.*, 39(12):1938–1953, 1993.

R.M.C. De Keyser and A.R. Van Cauwenberghe. Extended prediction self-adaptive control. In *7th IFAC Symposium on Identification and System Parameter Estimation*, pages 1255–1260, Oxford, 1985. Pergamon.

R.M.C. De Keyser, G.A. Van de Velde, and F.A.G. Dumortier. A comparative study of self-adaptive long-range predictive control methods. *Automatica*, 24 (2):149–163, 1988.

P.D. Khandalekar and J.B. Riggs. Nonlinear process model based control and optimization of a model IV FCC unit. *Comp. Chem. Engng.*, 19(11):1153–1168, 1995.

D. Kunii and O. Levenspiel. *Fluidization Engineering.* Butterworth, Stoneham, 2nd edition, 1991.

H.I. De Lasa, A. Errazu, E. Barreiro, and S. Solioz. Analysis of fluidized bed catalytic cracking regenerator models in an industrial scale unit. *Can. J. Chem. Eng.*, 59:549–553, 1981.

E. Lee and F.R.Jr. Groves. Mathematical model of the fluidized bed catalytic cracking plant. *Trans. Soc. Comput. Sim.*, 2:219–236, 1985.

J.H. Lee. Recent advances in model predictive control and other related areas. In *Chemical Process Control-CPC V*, Tahoe, California, 1996.

J.H. Lee, M. Morari, and C.E. Garcia. State-space interpretation of model predictive control. *Automatica*, 30:707–717, 1994.

L.S. Lee, Y.W. Chen, T.N. Huang, and W.Y. Pan. Four-lump kinetic model for fluid catalytic cracking process. *Can. J. Chem. Engng.*, 67:615–619, 1989.

S. Li, K.Y. Lim, and D.G. Fisher. A state space formulation for model predictive control. *AIChE. J.*, 35(2):241–249, 1989.

P. Lunström, J.H. Lee, M. Morari, and S. Skogestad. Limitations of dynamic matrix control. *Comp. Chem. Engng.*, 19(4):409–421, 1995.

J.M. Maciejowski. *Predictive Control.* Pearson Education, Harlow, England, 2002.

P. Malay. A Modified Integrated Dynamic Model of a Riser Type FCC Unit. Master's thesis, University of Saskatchewan, Saskatoon, Canada, 1998.

B.R. Maner, F.J. Doyle III, B.A. Ogunnaike, and R.K. Pearson. Nonlinear model predictive control of a simulated multivariable polymerization reactor using second-order Volterra models. *Automatica*, 32(9):1285–1301, 1996.

D.Q. Mayne. Nonlinear model predictive control: an assessment. In *Chemical Process Control-CPC V*, Tahoe, California, 1996.

D.Q. Mayne and H. Michalska. Receding horizon control of nonlinear systems. *IEEE Trans. Automat. Contr.*, AC35(7):814–824, 1990.

D.Q. Mayne, J.B. Rawlings, C.V. Rao, and P.O.M. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36:789–814, 2000.

R.C. McFarlane, R.C. Reinemann, J.F. Bartee, and C. Georgakis. Analysis of fluidized bed catalytic cracking regenerator models in an industrial scale unit. *Comp. Chem. Engng.*, 17:275–300, 1993.

H. Michalska and D.Q. Mayne. Robust receding horizon control of constrained nonlinear systems. *IEEE Trans. Automat. Contr.*, AC38(11):1623–1633, 1993.

J.J. Monge and C. Georgakis. Multivariable control of catalytic cracking processes. *Chem. Eng. Comm.*, 61:197–225, 1987.

M. Morari and J.H. Lee. Model predictive control: the good, the bad and the ugly. In Y. Arkun and W. H. Ray, editors, *Chemical Process Control - CPC IV*, pages 419–444, Amsterdam, 1991. Elsevier.

M. Morari and J.H. Lee. Model predictive control: past, present and future. *Comp. Chem. Engng.*, 23:667–682, 1999.

A.M. Morshedi, C.R. Cutler, and T.A. Skrovanek. Optimal solution of dynamic matrix control with linear programming techniques. pages 199–208. Proc. Am. Control. Conf., Boston, 1985.

K.R. Muske and J.B. Rawlings. Model predictive control with linear models. *AIChE. J.*, 39(2):262–287, 1993.

S.J. Qin and T.A. Badgwell. An overview of nonlinear model predictive control applications. In F. Allgöwer and A. Zheng, editors, *Non Linear Model Predictive Control*, pages 369–392. Birkhäuser, Basel, 2000.

S.J. Qin and T.A. Badgwell. An overview of industrial model control technology. In *Chemical Process Control - CPC V*, pages 232–255, Tahoe, California, 1996.

J.B. Rawlings and K.R. Muske. The stability of constrained receding control. *IEEE Transactions on Automatic Control*, 38(10):1512–1516, 1993.

J.B. Rawlings, E.S. Meadows, and K.R. Muske. Nonlinear model predictive control: A tutorial and survey. In *Advanced Control of Chemical Processes*, pages 203–224, Kyoto (Japan), 1994. IFAC.

J. Richalet, A. Rault, J.L. Testud, and J. Papon. Model predictive heuristic control: Applications to industrial processes. *Automatica*, 14:413–428, 1978.

N.L. Ricker. Model predictive control with state estimation. *Ind. Eng. Chem. Res.*, 29:374–382, 1990a.

N.L. Ricker. Model predictive control of processes with many inputs and outputs. *Control and System Dynamics*, 37:217–269, 1990b.

N.L. Ricker. Model predictive control: State of the art. In Y. Arkun and W. H. Ray, editors, *Chemical Process Control - CPC IV*, Amsterdam, 1991. Elsevier.

J.M. Martin Sanchez and J. Rodellar. *Adaptive Predictive Control.* Prentice Hall, Englewood Cliffs, New Jersey, 1996.

R. Shridar and D.J. Cooper. A novel tuning strategy for multivariable model predictive control. *ISA Transactions*, 36(4):273–280, 1998.

R. Soeterboek. *Predictive Control - A Unified Approach.* Prentice Hall, Englewood Cliffs, New Jersey, 1992.

J.G. Van Antwerp and R.D. Braatz. Model predictive control of large scale processes. *J. Proc. Cont.*, 10:1–8, 2000.

P. Vuthandam, H. Genceli, and M. Nikolaou. Performance bounds for robust quadratic dynamic matrix control with end condition. *AIChE J.*, 41(9): 2083–2097, 1995.

B.E. Ydstie. Extended horizon adaptive control. pages 911–915, Oxford, 1984. IFAC 9th World Congress Budapest Hungary, Pergamon.

Part V

Nonlinear Control

Chapter 17

Nonlinear Geometric Control

Most chemical processes display nonlinear behaviour, whether they are chemical or biological reactors, neutralization reactors and distillation columns. To control their operation around a steady-state point, a linear model, obtained by linearization or identification, is frequently used for designing the control law. In some cases of highly nonlinear behaviour or during the transient regimes such as startup and shutdown, in the case of batch or fed-batch processes, a linear controller is insufficient to guarantee the stability and performance. It is possible to use gain programming controllers, which need to test the controller tuning for all the operating domains crossed. Adaptive control itself is not always satisfactory and, in particular, can pose robustness problems. On the other hand, it is often possible to have at our disposal a nonlinear knowledge model of the process, which may be more or less accurate.

The theory of linear system control is much more developed than that of nonlinear system control, but the progress realized in the nonlinear domain is such that the designer of a control system nowadays can find a certain number of efficient tools, provided that he or she possesses nonlinear knowledge model of the process.

Among the different methods of nonlinear control, the theory related to differential geometry will be discussed in this chapter and the reader will be able to refer to the two applications of nonlinear geometric control concerning a chemical reactor and a biological reactor developed in Chap. 19. Other methods exist, but the proposed method is often the most employed, as it allows us to answer, in particular, questions of reachability, observability and feedback linearization. The problem is to algebraically transform the dynamics of a nonlinear system into partially or totally linear dynamics by a transformation acting on the states, which differs totally from the linear approximation of dynamics by calculation of the Jacobian.

17.1 Some Linear Notions Useful in Nonlinear Control

The theory of nonlinear control is an extension of the theory developed originally for linear systems (Wohnam, 1985). For this reason, a certain number of notions having common links with linear and nonlinear systems are discussed in a form which emphasizes the similarities between linear and nonlinear theory (Isidori, 1989; Khalil, 1996; Kravaris and Kantor, 1990a; Slotine and Li, 1991).

Initially, we consider single-input single-output systems in the most general form

$$\left\{ \begin{array}{ll} \dot{x} & = f(x,u) \\ y & = h(x) \end{array} \right. \tag{17.1}$$

where x is the state vector of dimension n, u the control input and y the controlled output. In fact, in chemical engineering, most systems are affine with respect to the input

$$\left\{ \begin{array}{ll} \dot{x} & = f(x) + g(x)\,u \\ y & = h(x) \end{array} \right. \tag{17.2}$$

$f(x)$ and $g(x)$ are respectively called vector fields of the dynamics and the control.

In the case of a linear system

$$\left\{ \begin{array}{ll} \dot{x} & = A\,x + B\,u \\ y & = C\,x \end{array} \right. \tag{17.3}$$

the vector fields are

$$f(x) = Ax \quad ; \quad g(x) = B \quad ; \quad h(x) = Cx \tag{17.4}$$

where A, B and C are matrices of respective dimensions $n \times n$, $n \times 1$ and $1 \times n$.

17.1.1 Influence of a Coordinate Change in Linear Control

The states x represent a set of coordinates which describe the time evolution of this system. The study of dynamic systems frequently requires a coordinate change. In the linear framework, a coordinate change is defined by a nonsingular similarity matrix T such that the new coordinates are

$$\xi = T\,x \tag{17.5}$$

The linear system (17.3) is thus transformed into

$$\left\{ \begin{array}{ll} \dot{\xi} & = \bar{A}\,\xi + \bar{B}\,u \\ y & = \bar{C}\,\xi \end{array} \right. \tag{17.6}$$

with

$$\bar{A} = T\,A\,T^{-1} \quad ; \quad \bar{B} = T\,B \quad ; \quad \bar{C} = C\,T^{-1} \tag{17.7}$$

The two linear systems (17.3) and (17.6) have the same transfer function

$$\frac{Y(s)}{U(s)} = C\,(s\,\boldsymbol{I} - A)^{-1}\,B = \frac{C\,\mathrm{Adj}(s\,\boldsymbol{I} - A)\,B}{\det(s\,\boldsymbol{I} - A)} \tag{17.8}$$

and represent two different realizations of the same input-output system. A series expansion of transfer function (17.8) gives

$$C\,(s\,\boldsymbol{I} - A)^{-1}\,B = \frac{C\,B}{s} + \frac{C\,A\,B}{s^2} + \frac{C\,A^2\,B}{s^3} + \ldots \tag{17.9}$$

where the quantities $C\,B$, $C\,A\,B$, $C\,A^2\,B$, ..., $C\,A^k\,B$, ... , are called Markov parameters of the system, which are sufficient to completely determine the system transfer function.

17.1.2 Relative Degree

The degree of the denominator $\det(s\,\boldsymbol{I} - A)$ of transfer function (17.8) is always equal to the system order n, but the degree of numerator $C\,\mathrm{Adj}(s\,I - A)\,B$ is included between 0 and $n - 1$. The relative degree of linear system (17.3), also called relative order or characteristic index, is equal to the difference between the degrees of denominator and numerator of transfer function (17.8).

Definition (Hirschorn, 1979): the relative degree of linear system (17.3) is defined by

$$\begin{array}{l} C\,A^k\,B = 0 \quad \text{for all:} \quad k < r - 1 \\ C\,A^{r-1}\,B \neq 0 \end{array} \tag{17.10}$$

The relative degree r is the smallest integer such that: $C\,A^{r-1}\,B \neq 0$. The system (17.3) will thus have a relative degree r such that

$$\begin{array}{lll} r = 1 & \text{if} & C\,B \neq 0 \\ r = 2 & \text{if} & C\,B = 0 \ \text{ and } \ C\,A\,B \neq 0 \\ r = 3 & \text{if} & C\,B = C\,A\,B = 0 \ \text{ and } \ C\,A^2\,B \neq 0, \ldots \end{array} \tag{17.11}$$

The relative degree, if it exists, is necessarily such that: $1 \leq r \leq n$.

The relative degree can also be considered from a different angle by taking into account the successive time derivatives of the output y

$$\begin{array}{ll} \dfrac{dy}{dt} & = C\,A\,x \\ & \vdots \\ \dfrac{d^{r-1}y}{dt^{r-1}} & = C\,A^{r-1}\,x \\ \dfrac{d^{r}y}{dt^{r}} & = C\,A^{r}\,x + C\,A^{r-1}\,B\,u \end{array} \tag{17.12}$$

by using the definition (17.10) of the relative degree. The relative degree is thus the smallest degree of differentiation of the output y which depends explicitly on the input u.

The design of a controller, e.g. in internal model control, often makes use of the inverse of the process transfer function. The inversion of a linear system consists of finding a state-space realization of the transfer function of the inverse of the model, that is

$$\frac{U(s)}{Y(s)} = \frac{\det(s\,\boldsymbol{I} - A)}{C\,\mathrm{Adj}(s\,\boldsymbol{I} - A)\,B} \tag{17.13}$$

The dynamic system

$$\left\{\begin{array}{lll} \dot{z} & = (A - \dfrac{B\,C\,A^r}{C\,A^{r-1}\,B})\,z + \dfrac{B}{C\,A^{r-1}\,B}\dfrac{d^r y}{dt^r} \\ u & = \dfrac{1}{C\,A^{r-1}\,B}\dfrac{d^r y}{dt^r} - \dfrac{C\,A^r}{C\,A^{r-1}\,B}\,z \end{array}\right. \tag{17.14}$$

is a state-space realization of the inverse (17.13) of system (17.3) assumed to be of relative degree r (Kravaris and Kantor, 1990a). However, this is not a minimal order realization of the inverse. Indeed, the order of dynamic system (17.14) is n, whereas the degree of transfer function (17.13) is $n - r$. This can also be verified by calculating the transfer function of (17.14): there are r pole simplifications at the origin.

17.1.3 Normal Form and Relative Degree

The vectors C, $C\,A$, ..., $C\,A^{r-1}$ are linearly independent. We assume $b_1 \neq 0$.

In these conditions, the transformation

$$\begin{array}{lll} \xi_1 & = C\,x \\ & \vdots \\ \xi_r & = C\,A^{r-1}\,x \\ \xi_{r+1} & = x_{r+1} - \dfrac{b_{r+1}}{b_1}\,x_1 \\ & \vdots \\ \xi_n & = x_n - \dfrac{b_n}{b_1}\,x_1 \end{array} \tag{17.15}$$

is invertible and transforms system (17.3) into the following system

$$\begin{array}{lll} \dot{\xi}_1 & = \xi_2 \\ & \vdots \\ \dot{\xi}_{r-1} & = \xi_r \\ \dot{\xi}_r & = \tilde{a}_1 \begin{bmatrix} \xi_{r+1} \\ \vdots \\ \xi_n \end{bmatrix} + \gamma_1 \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_r \end{bmatrix} + \beta\,u \\ \begin{bmatrix} \dot{\xi}_{r+1} \\ \vdots \\ \dot{\xi}_n \end{bmatrix} & = \tilde{A} \begin{bmatrix} \xi_{r+1} \\ \vdots \\ \xi_n \end{bmatrix} + \Gamma \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_r \end{bmatrix} \\ y & = \xi_1 \end{array} \tag{17.16}$$

where $\tilde{A}$, Γ, $\tilde{a}_1$, γ_1 are matrices of respective dimensions $(n-r)\times(n-r)$, $(n-r)\times r$, $1\times(n-r)$, $1\times r$, and β is a nonzero scalar.

The dynamic system (17.16) is a different realization of transfer function (17.8) and constitutes a normal form of any linear system of relative degree r. A remarkable property of the realization (17.16) is that the $(n-r)$ eigenvalues of matrix $\tilde{A}$ are the zeros of transfer function (17.8) by noticing that the transfer function of system (17.16) is equal to

$$\frac{Y(s)}{U(s)} = \frac{\beta \det(s\boldsymbol{I} - \tilde{A})}{\det(s\boldsymbol{I} - \tilde{A})\left(s^r - \gamma_1 \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{r-1} \end{bmatrix}\right) - \tilde{a}_1 \operatorname{Adj}(s\boldsymbol{I} - \tilde{A})\Gamma \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{r-1} \end{bmatrix}} \tag{17.17}$$

From the normal form (17.16), it appears that the relative degree r is the number of integrators that the input u must cross before affecting the output $y = \xi_1$. The normal form also allows us to obtain a minimal order realization of the inverse system (17.13). By using the derivatives of the output, the system (17.16) becomes

$$\begin{aligned}
\dot{\xi}_1 &= \dot{y} = \xi_2 \\
&\vdots \\
\dot{\xi}_{r-1} &= \frac{d^{r-1}y}{dt^{r-1}} = \xi_r \\
\dot{\xi}_r &= \frac{d^r y}{dt^r} = \tilde{a}_1 \begin{bmatrix} \xi_{r+1} \\ \vdots \\ \xi_n \end{bmatrix} + \gamma_1 \begin{bmatrix} y \\ \dfrac{dy}{dt} \\ \vdots \\ \dfrac{d^{r-1}y}{dt^{r-1}} \end{bmatrix} + \beta u \\
\begin{bmatrix} \dot{\xi}_{r+1} \\ \vdots \\ \dot{\xi}_n \end{bmatrix} &= \tilde{A} \begin{bmatrix} \xi_{r+1} \\ \vdots \\ \xi_n \end{bmatrix} + \Gamma \begin{bmatrix} y \\ \dfrac{dy}{dt} \\ \vdots \\ \dfrac{d^{r-1}y}{dt^{r-1}} \end{bmatrix}
\end{aligned} \tag{17.18}$$

which gives the control

$$u = \frac{1}{\beta}\left\{ \frac{d^r y}{dt^r} - \tilde{a}_1 \begin{bmatrix} \xi_{r+1} \\ \vdots \\ \xi_n \end{bmatrix} - \gamma_1 \begin{bmatrix} y \\ \dfrac{dy}{dt} \\ \vdots \\ \dfrac{d^{r-1}y}{dt^{r-1}} \end{bmatrix} \right\} \tag{17.19}$$

In this form, only the states $(\xi_{r+1}, \ldots, \xi_n)$ intervene, and they are denoted by $(z_1, \ldots, z_{n-r})$ by subscript translation. It is possible to deduce a minimal realization of the inverse system (17.13)

$$\begin{aligned} \begin{bmatrix} \dot{z}_1 \\ \vdots \\ \dot{z}_{n-r} \end{bmatrix} &= \tilde{A} \begin{bmatrix} z_1 \\ \vdots \\ z_{n-r} \end{bmatrix} + \Gamma \begin{bmatrix} y \\ \dfrac{dy}{dt} \\ \vdots \\ d^{r-1}y/dt^{r-1} \end{bmatrix} \\ u &= \frac{1}{\beta} \left\{ \frac{d^r y}{dt^r} - \tilde{a}_1 \begin{bmatrix} z_1 \\ \vdots \\ z_{n-r} \end{bmatrix} - \gamma_1 \begin{bmatrix} y \\ \dfrac{dy}{dt} \\ \vdots \\ d^{r-1}y/dt^{r-1} \end{bmatrix} \right\} \end{aligned} \tag{17.20}$$

17.1.4 Zero Dynamics

The zeros of a linear system are the roots of the numerator polynomial of transfer function (17.8) and can be considered as the poles of its inverse (17.13). As the realization (17.20) of the inverse is minimal, the zeros of transfer function (17.8) are the roots of $\det(sI - \tilde{A}) = 0$. The zero dynamics of a system is defined as the dynamics of its inverse. Thus, the dynamics of the last $(n-r)$ equations of the normal form (17.16) completely determines the zeros of the system, i.e.

$$\begin{bmatrix} \dot{z}_1 \\ \vdots \\ \dot{z}_{n-r} \end{bmatrix} = \tilde{A} \begin{bmatrix} z_1 \\ \vdots \\ z_{n-r} \end{bmatrix} + \Gamma \begin{bmatrix} y \\ dy/dt \\ \vdots \\ d^{r-1}y/dt^{r-1} \end{bmatrix} \tag{17.21}$$

The dynamic system (17.21) written as

$$\begin{bmatrix} \dot{z}_1 \\ \vdots \\ \dot{z}_{n-r} \end{bmatrix} = \tilde{A} \begin{bmatrix} z_1 \\ \vdots \\ z_{n-r} \end{bmatrix} + \Gamma \begin{bmatrix} u_1 \\ \vdots \\ u_r \end{bmatrix} \tag{17.22}$$

is called the forced dynamics of system (17.3), while the dynamic system

$$\begin{bmatrix} \dot{z}_1 \\ \vdots \\ \dot{z}_{n-r} \end{bmatrix} = \tilde{A} \begin{bmatrix} z_1 \\ \vdots \\ z_{n-r} \end{bmatrix} \tag{17.23}$$

is called the unforced dynamics or internal dynamics of system (17.3), or zero dynamics of system (17.3). Actually, the system (17.23) completely determines the zero dynamics. The stability of the internal dynamics is thus determined by the positions of the zeros.

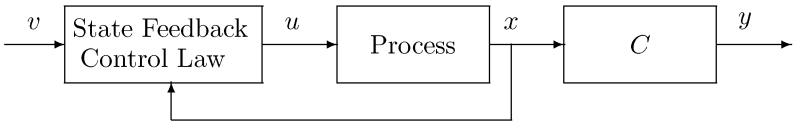


Figure 17.1: Static state feedback control

17.1.5 Static State Feedback

Consider the linear system (17.3) subjected to the static state feedback control

$$u = \lambda\, v - K x \tag{17.24}$$

where v is an external scalar input (Fig. 17.1), λ a scalar gain for external input precompensation and K a state feedback gain vector of dimension n.

The resulting closed-loop system is

$$\left\{ \begin{array}{ll} \dot{x} & = (A - B\,K)x + \lambda\, B\, v \\ y & = C\, x \end{array} \right. \tag{17.25}$$

The static state feedback (17.24) presents the following properties:

– It preserves the linearity, the relative degree and the zeros of the system.
– It modifies the poles of the system (which is the aim of the state feedback).

A static state feedback control law can be defined for the system (17.3) of relative degree r. The order r derivative of the output y of system (17.3) is equal to

$$y^{(r)} = C\, A^r x + C\, A^{r-1}\, B\, u \tag{17.26}$$

as $C\, A^{r-1}\, B \neq 0$. The static state feedback control law results

$$u = \frac{v - C\, A^r\, x}{C\, A^{r-1}\, B} \tag{17.27}$$

with $v = y^{(r)}$. Note that the static state feedback control law (17.27) is a minimal order realization of the inverse of system (17.3). The linear system (17.3) subjected to the static state feedback (17.27) becomes

$$\left\{ \begin{array}{ll} \dot{x} & = (A - B\, \dfrac{C\, A^r}{C\, A^{r-1}\, B})\, x + B\, \dfrac{v}{C\, A^{r-1}\, B} \\ y & = C\, x \end{array} \right. \tag{17.28}$$

17.1.6 Pole-Placement by Static State Feedback

Consider a linear system in its controllable canonical form. The associated dynamic system is

$$\begin{array}{ll} \dot{x}_1 & = x_2 \\ & \vdots \\ \dot{x}_{n-1} & = x_n \\ \dot{x}_n & = -\tilde{\alpha}_n x_1 - \tilde{\alpha}_{n-1} x_2 - \ldots - \tilde{\alpha}_1 x_n + \tilde{b} u \end{array} \tag{17.29}$$

and its characteristic polynomial is

$$s^n + \tilde{\alpha}_1 s^{n-1} + \ldots + \tilde{\alpha}_{n-1} s + \tilde{\alpha}_n \tag{17.30}$$

Suppose that we desire a static state feedback that places the poles such that the characteristic polynomial becomes

$$s^n + \alpha_1 s^{n-1} + \ldots + \alpha_{n-1} s + \alpha_n \tag{17.31}$$

The static state feedback realizing this pole-placement is equal to

$$u = \frac{1}{\tilde{b}} \left(v - [\alpha_n - \tilde{\alpha}_n \quad \ldots \quad \alpha_1 - \tilde{\alpha}_1] \, x \right) \tag{17.32}$$

For any controllable linear system such as

$$\dot{x} = A \, x + B \, u \tag{17.33}$$

the previous result is easily generalized by considering the transformation T such that

$$\xi = T \, x = \begin{bmatrix} Q \\ Q \, A \\ \vdots \\ Q \, A^{n-1} \end{bmatrix} x \tag{17.34}$$

where Q is a row vector satisfying

$$\begin{aligned} &Q \, B = 0 \\ &Q \, A \, B = 0 \\ &\vdots \\ &Q \, A^{n-2} \, B = 0 \\ &Q \, A^{n-1} \, B \neq 0 \end{aligned} \tag{17.35}$$

This transformation transforms the system (17.33) into

$$\begin{aligned} \dot{\xi}_1 \quad &= \xi_2 \\ &\vdots \\ \dot{\xi}_{n-1} \quad &= \xi_n \\ \dot{\xi}_n \quad &= Q \, A^n \, T^{-1} \, \xi + Q \, A^{n-1} \, B \, u \end{aligned} \tag{17.36}$$

From Eq. (17.32), the static state feedback realizing the pole-placement (17.31) is thus equal to

$$u = \frac{1}{Q \, A^{n-1} \, B} \left(v - ([\alpha_n \; \alpha_{n-1} \; \ldots \; \alpha_1] + Q \, A^n \, T^{-1}) \, T \, x \right) \tag{17.37}$$

which can be transformed, according to Ackermann's formula

$$u = \frac{1}{Q \, A^{n-1} \, B} \left(v - (Q \, A^n + \alpha_1 \, Q \, A^{n-1} + \ldots + \alpha_{n-1} \, Q \, A + \alpha_n \, Q) \, x \right) \tag{17.38}$$

Among the possible vectors Q, the following vector is suitable

$$Q = \text{last row of } [B \; AB \; A^{n-1}B]^{-1} \tag{17.39}$$

with, moreover, $QA^{n-1}B = 1$.

17.1.7 Input-Output Pole-Placement

Consider the linear system (17.3) assumed of relative degree r. We seek a state feedback which imposes a given transfer function for the closed-loop system. The static state feedback preserves the relative degree of the system, thus the simplest closed-loop transfer function can be written as

$$\frac{Y(s)}{V(s)} = \frac{1}{s^r + \beta_1 s^{r-1} + \ldots + \beta_{r-1} s + \beta_r} \tag{17.40}$$

The static state feedback which, when applied to the system (17.3), gives transfer function (17.40), is equal to

$$u = \frac{1}{CA^{r-1}B} \left[v - \left(CA^r + \beta_1 CA^{r-1} + \ldots + \beta_{r-1} CA + \beta_r C \right) x \right] \tag{17.41}$$

By identification, the general static state feedback control law (17.24) gives

$$\begin{aligned} \lambda &= \frac{1}{CA^{r-1}B} \\ K &= \frac{1}{CA^{r-1}B} \left(CA^r + \beta_1 CA^{r-1} + \ldots + \beta_{r-1} CA + \beta_r C \right) \end{aligned} \tag{17.42}$$

The corresponding closed-loop characteristic polynomial is equal to

$$\det(sI - A + BK) = \frac{1}{CA^{r-1}B} \, C \, \mathrm{Adj}(sI - A) \, B \left(s^r + \beta_1 s^{r-1} + \ldots + \beta_{r-1} s + \beta_r \right) \tag{17.43}$$

According to this expression, we notice that the state feedback (17.41) places the closed-loop poles at the process zeros and the desired poles specified by Eq. (17.40). The closed-loop system can thus only be stable (internal stability of the states) if its zeros have a negative real part, thus if the system is minimum-phase.

17.2 Monovariable Nonlinear Control

17.2.1 Some Notions of Differential Geometry

Several textbooks have been published recently in the domain of nonlinear control, in particular the following: Fossard and Normand-Cyrot (1993); Isidori (1989, 1995); Nijmeijer and VanderSchaft (1990); Slotine and Li (1991); Vidyasagar (1993). In order to master the concepts which will be used in this chapter, some elementary notions of differential geometry must be introduced.

We consider nonlinear systems, affine with respect to the input, such as

$$\left\{ \begin{aligned} \dot{x} &= f(x) + g(x)\, u \\ y &= h(x) \end{aligned} \right. \tag{17.44}$$

where $f(x)$, $g(x)$ are smooth mappings and $h(x)$ a smooth function[1]. $f(x)$ and $g(x)$ are frequently called vector fields.

[1] A function or a mapping is said to be smooth if it possesses continuous partial derivatives of any order; mathematically the function is C^∞.

Given two real vectors V and W, the scalar product of the two vectors is denoted by

$$< V, W >= V^T W = \sum_{i=1}^{n} v_i \, w_i \tag{17.45}$$

We use the following notation for the gradient vector (taken as a row vector) of a function λ

$$D\lambda(x) = \frac{\partial \lambda}{\partial x} = \left[\begin{array}{ccc} \frac{\partial \lambda}{\partial x_1} & \dots & \frac{\partial \lambda}{\partial x_n} \end{array} \right] \tag{17.46}$$

The gradient (as a row vector) of a function λ is thus the Jacobian matrix of λ.

The vector fields $g_1(x), \dots, g_m(x)$ are linearly independent if, for all x, the vectors $g_1(x), \dots, g_m(x)$ are linearly independent.

The scalar fields $z_1(x), \dots, z_m(x)$ are linearly independent if, for all x, their gradients $Dz_1(x), \dots, Dz_m(x)$ are linearly independent vector fields.

The derivative of a function $\lambda(x)$ in the direction of the field f (directional derivative) is defined by

$$L_f \lambda(x) = \sum_{i=1}^{n} \frac{\partial \lambda}{\partial x_i} \, f_i(x) =< \frac{\partial \lambda}{\partial x}, f(x) > \tag{17.47}$$

This derivative is called the Lie derivative and plays a very important role in nonlinear control. As a matter of fact, for the system (17.2)

$$\begin{array}{ll} \frac{dy}{dt} & = \sum_{i=1}^{n} \frac{\partial h}{\partial x_i} \frac{dx_i}{dt} \\ & = \sum_{i=1}^{n} \frac{\partial h}{\partial x_i} \left(f_i(x) + g_i(x) \, u \right) \\ & = L_f h(x) + L_g h(x) \, u \end{array} \tag{17.48}$$

thus the time derivative of the output is simply expressed with respect to the Lie derivatives. The Lie derivative $L_f \lambda(x)$ is the derivative of λ along the integral curves[2]of the vector field f.

It is possible to realize successive differentiations, such as the differentiation of λ in the direction of f, then in the direction of g, that is

$$L_g L_f \lambda(x) = \frac{\partial L_f \lambda}{\partial x} \, g(x) \tag{17.49}$$

or further, to differentiate λ, k times in the direction of f

$$L_f^k \lambda(x) = \frac{\partial L_f^{k-1} \lambda}{\partial x} \, f(x) \qquad \text{with: } L_f^0 \lambda(x) = \lambda(x) \tag{17.50}$$

[2]Given the state-space system

$$\dot{x}(t) = f(x(t)) \qquad ; \quad x(0) = x^\circ$$

the solution $x(t)$ passing by x° defines a curve called the integral curve.

The Lie bracket is defined by

$$[f,g](x) = \frac{\partial g}{\partial x} f(x) - \frac{\partial f}{\partial x} g(x) \tag{17.51}$$

where $\partial f/\partial x$ is the Jacobian matrix of f equal to (same for $\partial g/\partial x$, the Jacobian matrix of g)

$$Df(x) = \frac{\partial f}{\partial x} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{17.52}$$

It is possible to repeat the operation on the Lie bracket of g by iterating on f, but in this case, rather than denoting by $[f,[f,\ldots,[f,g]]]$, the following notation is used

$$\mathrm{ad}_f^k g(x) = [f, \mathrm{ad}_f^{k-1} g](x) \qquad \text{for } k > 1 \quad \text{with: } \mathrm{ad}_f^0 g(x) = g(x) \tag{17.53}$$

The Lie bracket is a bilinear, skew-symmetric mapping and satisfies the Jacobi identity

$$[f,[g,p]] + [g,[p,f]] + [p,[f,g]] = 0 \tag{17.54}$$

where f, g, p are vector fields.

17.2.2 Relative Degree of a Monovariable Nonlinear System

Consider the single-input single-output nonlinear system (17.2). The relative degree, or relative order, or characteristic index, is defined (Hirschorn, 1979) by:
Definition: the relative degree of the nonlinear system (17.2) over a domain U is the smallest integer r for which

$$L_g L_f^{r-1} h(x) \neq 0 \qquad \text{for all } x \text{ in } U \tag{17.55}$$

This relation is consistent with the definition of the relative degree for linear systems by noticing that, for a linear system, Eq. (17.55) becomes

$$L_g L_f^{r-1} h(x) = C A^{r-1} B \neq 0 \tag{17.56}$$

The nonlinear system (17.2) will thus admit a relative degree r equal to

$$\begin{array}{ll} r = 1 & \text{if } L_g h(x) \neq 0 \\ r = 2 & \text{if } L_g h(x) = 0 \text{ and } L_g L_f h(x) \neq 0 \\ r = 3 & \text{if } L_g h(x) = L_g L_f h(x) = 0 \text{ and } L_g L_f^2 h(x) \neq 0 \\ \ldots & \end{array} \tag{17.57}$$

Isidori (1995) defines the relative degree at a point x° if the definition (17.55) is valid at any point x of the neighbourhood U; however, the first Lie derivative $L_g h(x)$ of the sequence $L_g L_f^{k-1} h(x)$ can be accidentally zero at x°. In this case, the relative degree cannot be defined strictly at x°, but will be defined in the neighbourhood U (notion of dense open subset). This point will be present in the following, although it will not be systematically recalled.

The condition (17.55) can be transformed, by use of the Leibnitz formula (Kravaris and Kantor, 1990a), and formulated as: the relative degree of the nonlinear system (17.2) is the smallest integer r such that

$$L_{\mathrm{ad}_f^{r-1} g} h(x) \neq 0 \tag{17.58}$$

As in the linear case, the interpretation of the relative degree r can be obtained from the time derivatives of the output y

$$\begin{array}{lll} \dfrac{dy}{dt} & = L_f h(x) + L_g h(x)\, u & \\ & = L_f h(x) & \text{if } \ 1 < r \\ & \vdots & \\ \dfrac{d^k y}{dt^k} & = L_f^k h(x) + L_g L_f^{k-1} h(x)\, u & \\ & = L_f^k h(x) & \text{if } k < r \\ \dfrac{d^r y}{dt^r} & = L_f^r h(x) + L_g L_f^{r-1} h(x)\, u & \text{as } \ L_g L_f^{r-1} h(x) \neq 0 \end{array} \tag{17.59}$$

The relative degree is equal to the number of differentiations of the output y that are necessary to make the input u explicitly appear.

If we obtain

$$L_g L_f^k h(x) = 0 \qquad \text{for all } k \qquad , \quad \text{for all } x \text{ in } U \tag{17.60}$$

the relative degree cannot be defined in the neighbourhood of x° and the output is not touched by the input u.

The row vectors $Dh(x), DL_f h(x), \ldots, DL_f^{r-1} h(x)$ are linearly independent (Isidori, 1995). This can be proved by showing that the matrix

$$\begin{bmatrix} Dh(x) \\ DL_f h(x) \\ \vdots \\ DL_f^{r-1} h(x) \end{bmatrix} \begin{bmatrix} g(x) & \mathrm{ad}_f g(x) & \ldots & \mathrm{ad}_f^{r-1} g(x) \end{bmatrix} \tag{17.61}$$

has rank r. This is an important point, as the r functions $h(x)$, $L_f h(x)$, ..., $L_f^{r-1} h(x)$ can form a new set of coordinates in the neighbourhood of point x°.

17.2.3 Frobenius Theorem

The Frobenius theorem is related to the solving of a first-order partial differential equation:

a) First, suppose that we have d smooth vector fields $f_i(x)$, defined on Ω°, which span a distribution[3] Δ, denoted by

$$\Delta = \text{span}\{f_1(x), \ldots, f_d(x)\} \tag{17.62}$$

In the same neighbourhood, the codistribution Ω, of dimension $n-d$ is spanned by $n-d$ covector fields: $\omega_1, \ldots, \omega_{n-d}$ such that

$$< \omega_j(x), f_i(x) >= 0 \qquad \forall 1 \leq i \leq d\,,\ 1 \leq j \leq n-d \tag{17.63}$$

Because of this property, we denote the codistribution: $\Omega = \Delta^\perp$, and ω_j is the solution of the equation

$$\omega_j(x) F(x) = 0 \tag{17.64}$$

where $F(x)$ is the matrix of dimension $n \times d$, of rank d, equal to

$$F(x) = \left[\ f_1(x) \quad \ldots \quad f_d(x)\ \right] \tag{17.65}$$

The row vectors ω_j form a basis of the space of the solutions of Eq. (17.64).
b) We assume that we are looking for solutions such that

$$\omega_j = \frac{\partial \lambda_j}{\partial x} \tag{17.66}$$

corresponding to smooth functions λ_j, i.e. we are looking for $n-d$ independent solutions (the row vectors $\partial\lambda_j/\partial x_1, \ldots, \partial\lambda_j/\partial x_{n-d}$ are independent) of the following differential equation

$$\frac{\partial \lambda_j}{\partial x} F(x) = \frac{\partial \lambda_j}{\partial x} \left[\ f_1(x) \quad \ldots \quad f_d(x)\ \right] = 0 \tag{17.67}$$

c) We seek the condition of existence of $n-d$ independent solutions of differential Eq. (17.67), which is equivalent to seeking the integrability of the distribution Δ: a distribution of dimension d, defined on an open domain U of $\mathbb{R}^n$, is completely integrable if, for all point x° of U, there exist $n-d$ smooth functions, having real values, defined on a neighbourhood of x°, such that

$$\text{span}\{\frac{\partial \lambda_1}{\partial x}, \ldots, \frac{\partial \lambda_{n-d}}{\partial x}\} = \Delta^\perp \tag{17.68}$$

The condition of existence is provided by the Frobenius theorem.
d) Frobenius theorem: a distribution is nonsingular if and only if it is involutive[4].

In the case where F is reduced to only a vector field f_1, Eq. (17.67) can be geometrically interpreted as:

[3] The smooth vector fields $f_1(x), \ldots, f_d(x)$ span at a point x of U a vector space dependent on x that can be denoted by $\Delta(x)$. The mapping assigning this vector space to any point x is called a smooth distribution.

[4] A distribution Δ is involutive if the Lie bracket of any couple of vector fields belonging to Δ belongs to Δ

$$f_1 \text{ and } f_2 \in \Delta \Longrightarrow [f_1, f_2] \in \Delta \Longleftrightarrow$$
$$[f_i, f_j](x) = \sum_{k=1}^{m} \alpha_{ijk} f_k(x) \qquad \forall i, j$$

– The gradient of λ is orthogonal to f_1.
– The vector f_1 is tangent to the surface $\lambda =$ constant passing by this point.
– The integral curve of f_1 passing by this point is entirely on the surface $\lambda =$ constant.

17.2.4 Coordinates Change

A function Φ of $\mathbb{R}^n$ in $\mathbb{R}^n$, defined in a domain U, is called a diffeomorphism if it is smooth and if its inverse Φ^{-1} exists and is smooth. If the domain U is the whole space, the diffeomorphism is global; otherwise, it is local. The diffeomorphism is thus a nonlinear coordinate change that possesses the properties previously listed.

Consider a function Φ defined in a domain U of $\mathbb{R}^n$. If and only if the Jacobian matrix $\partial\Phi/\partial x$ is nonsingular at x° belonging to Ω, $\Phi(x)$ defines a local diffeomorphism on a sub-domain Ω° of Ω.

A diffeomorphism allows us to transform a nonlinear system into another nonlinear system defined with regard to new states.

Given the single-input single-output nonlinear system (17.2) of relative degree r at x°, set

$$\begin{array}{ll} \phi_1(x) & = h(x) \\ \phi_2(x) & = L_f h(x) \\ & \vdots \\ \phi_r(x) & = L_f^{r-1} h(x) \end{array} \tag{17.69}$$

If $r < n$, it is possible to find $n - r$ functions $\phi_{r+1}(x), \ldots, \phi_n(x)$ such that the mapping

$$\Phi(x) = \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_n(x) \end{bmatrix} \tag{17.70}$$

has its Jacobian matrix nonsingular and thus constitutes a possible coordinate change at x°.

The value of the functions $\phi_{r+1}(x), \ldots, \phi_n(x)$ at x° has no importance and these functions can be chosen such that

$$< D\phi_i(x), g(x) >= L_g \phi_i(x) = 0 \quad \text{for all } r+1 \leq i \leq n \quad \text{for all } x \text{ in } \Omega \tag{17.71}$$

The demonstration makes use of the Frobenius theorem (Isidori, 1995).

17.2.5 Normal Form

The nonlinear system (17.2) can be described in the new coordinates $z_i = [y, \dot{y}, \ldots, y^{(r-1)}] = \phi_i(x), (i = 1, \ldots, n)$ from the relations (17.70)

$$\begin{aligned}
\frac{dz_1}{dt} &= \frac{\partial \phi_1}{\partial x}\frac{dx}{dt} = \frac{\partial h}{\partial x}\frac{dx}{dt} = L_f h(x(t)) = \phi_2(x(t)) = z_2(t) \\
&\vdots \\
\frac{dz_{r-1}}{dt} &= \frac{\partial \phi_{r-1}}{\partial x}\frac{dx}{dt} = \frac{\partial L_f^{r-2} h}{\partial x}\frac{dx}{dt} = L_f^{r-1} h(x(t)) = \phi_r(x(t)) = z_r(t) \\
\frac{dz_r}{dt} &= \frac{\partial \phi_r}{\partial x}\frac{dx}{dt} = \frac{\partial L_f^{r-1} h}{\partial x}\frac{dx}{dt} = L_f^r h(x(t)) + L_g L_f^{r-1} h(x(t))\, u(t)
\end{aligned} \tag{17.72}$$

The expression of $\dot{z}_r(t)$ must be transformed with respect to $z(t)$ by using the relation $x(t) = \Phi^{-1}(z(t))$, which gives

$$\begin{aligned}
\frac{dz_r}{dt} &= L_f^r h(\Phi^{-1}(z(t))) + L_g L_f^{r-1} h(\Phi^{-1}(z(t)))\, u(t) \\
&= b(z(t)) + a(z(t))\, u(t)
\end{aligned} \tag{17.73}$$

by setting

$$a(z(t)) = L_g L_f^{r-1} h(\Phi^{-1}(z(t))) \quad ; \quad b(z(t)) = L_f^r h(\Phi^{-1}(z(t))) \tag{17.74}$$

and by noting that, by definition of the relative degree, at $z^\circ = \Phi(x^\circ)$, we have: $a(z^\circ) \neq 0$.

It is possible to choose the following coordinates z_i, $r < i \leq n$, according to Eq. (17.71) so that $L_g \phi_i(x) = 0$, which gives

$$\begin{aligned}
\frac{dz_i}{dt} &= \frac{\partial \phi_i}{\partial x}\frac{dx}{dt} = \frac{\partial \phi_i}{\partial x}\left(f(x(t)) + g(x(t))\, u(t)\right) \quad ; \quad r < i \leq n \\
&= L_f \phi_i(x(t)) + L_g \phi_i(x(t))\, u(t)) = L_f \phi_i(x(t)) \\
&= L_f \phi_i(\Phi^{-1}(z(t)))
\end{aligned} \tag{17.75}$$

We set

$$q_i(z(t)) = L_f \phi_i(\Phi^{-1}(z(t))) \quad ; \quad r < i \leq n \tag{17.76}$$

By again considering all the equations, we obtain the normal form

$$\begin{aligned}
\dot{z}_1 &= z_2 \\
&\vdots \\
\dot{z}_{r-1} &= z_r \\
\dot{z}_r &= b(z) + a(z)\, u(t) \\
\dot{z}_{r+1} &= q_{r+1}(z) \\
\dot{z}_n &= q_n(z)
\end{aligned} \tag{17.77}$$

to which we must add the equation of the output

$$y = h(x) = z_1 \tag{17.78}$$

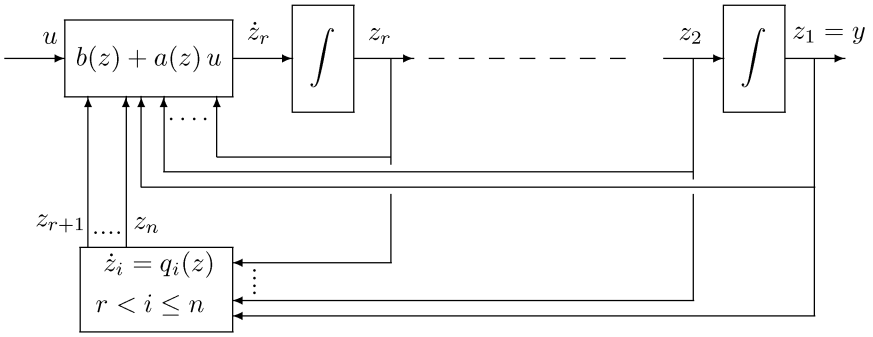


Figure 17.2: Description of the normal form

This result can be symbolized in the block diagram (Fig. 17.2) by displaying the chain of r integrators necessary to pass from the control input to the output.

The condition $L_g\phi_i(x) = 0$, which is fundamental for seeking the functions ϕ_i, $r < i \leq n$, can be difficult to fill, as this condition corresponds to the solving of a system of $n-r$ partial differential equations. To define a coordinate change, it can be sufficient to find these functions so that the matrix Φ is simply nonsingular.

17.2.6 Controllability and Observability

Consider the nonlinear system

$$\dot{x} = f(x) + g(x)\, u \tag{17.79}$$

defined in a domain U.

The system (17.79) is controllable if, given an initial state x°, whatever two points x_1 and x_2 of U, there exists an admissible input u such that the system passes from the state x_1 to the state x_2 in a finite time T.

The controllability of this nonlinear system can be studied by proceeding to a linearization of the system

$$\dot{z} = \frac{\partial f}{\partial x} z + g(x)\, v \tag{17.80}$$

and by studying the controllability matrix. However, this approach is not always satisfactory; indeed, a nonlinear system can be controllable whereas its linear approximation is not. It is necessary to introduce the notion of reachability (Isidori, 1995; Nijmeijer and VanderSchaft, 1990).

For observability that also requires complex topological notions, both previous books are recommended. Observability can be defined in an approached manner by:
A system is observable if, given two different initial conditions $x(0)$ and $x'(0)$, there exists a control input $u(t)$ defined in $[0, T]$ such that the corresponding

outputs $y(x, u, t)$ and $y'(x, u, t)$ are not identically equal in $[0, T]$. The input $u(t)$ distinguishes the initial conditions $x(0)$ and $x'(0)$ in $[0, T]$. If $u(t)$ distinguishes any pair (x, x') in $[0, T]$, the input $u(t)$ is universal.

17.2.7 Principle of Feedback Linearization

The control law depends on the state values, which are assumed to be known. In the case where all the states are not known, it is necessary to couple a state estimator, called an observer, to the control system. When the state feedback control law depends only on the values of the states x and the external input v, it is a static state feedback. If the control law corresponds to the output of a dynamic system, itself depending on the states x and on the external input v, it is a dynamic state feedback.

Consider the single-input single-output nonlinear system affine with respect to the input

$$\dot{x} = f(x) + g(x)\, u \tag{17.81}$$

defined in a neighbourhood U of x° and such that: $f(x^\circ) = 0$. The problem of feedback linearization is to find smooth functions p and q with $q(x^\circ) \neq 0$, and a diffeomorphism Φ with $\Phi(x^\circ) = 0$ such that by defining:

- an external input $v = p(x) + q(x)\, u$,
- the transformed variables $z = \Phi(x)$,

the resulting system is linear as

$$\dot{z} = A\, z + B\, v \tag{17.82}$$

where the pair (A, B) is controllable. The new state z is called a linearizing state and the control law is a linearizing control law.

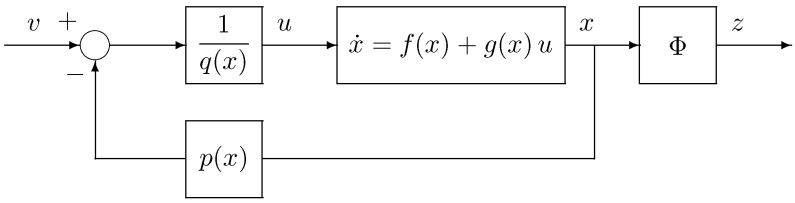


Figure 17.3: Linearizing feedback

With regard to the state x, the control law (Fig. 17.3) is

$$u(t) = \frac{-p(x)}{q(x)} + \frac{v}{q(x)} = \alpha(x) + \beta(x)\, v \tag{17.83}$$

17.2.8 Exact Input-State Linearization for a System of Relative Degree Equal to n

This linearization is often called exact linearization (Isidori, 1995). To start, consider the system (17.81) of relative degree $r = n$, thus equal to the dimension of the state vector, at a point x°.

The coordinate change necessary to get the normal form is then

$$\Phi(x) = \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_n \end{bmatrix} = \begin{bmatrix} h(x) \\ \vdots \\ L_f^{n-1} h(x) \end{bmatrix} \tag{17.84}$$

and the normal form is

$$\begin{aligned} \dot{z}_1 &= z_2 \\ &\vdots \\ \dot{z}_{n-1} &= z_n \\ \dot{z}_n &= b(z) + a(z)\, u(t) \end{aligned} \tag{17.85}$$

with $a(z^\circ) \neq 0$ because of the definition of the relative degree.

Figure 17.4: Exactly linearized system by static state feedback

If we choose the state feedback control law

$$u(t) = -\frac{b(z)}{a(z)} + \frac{v}{a(z)} \tag{17.86}$$

we obtain the resulting closed-loop system (Fig. 17.3)

$$\begin{aligned} \dot{z}_1 &= z_2 \\ &\vdots \\ \dot{z}_{n-1} &= z_n \\ \dot{z}_n &= v \end{aligned} \tag{17.87}$$

which is a linear and controllable system expressed in the Brunovsky canonical form

$$\dot{z} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \dots & & \dots & 0 \end{bmatrix} z + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix} v \tag{17.88}$$

To obtain the control law (17.86), a coordinate change and a state feedback which can be exchanged are used. If we first use the state feedback and then the coordinate change, we obtain the following control law

$$\begin{aligned} u(t) &= -\frac{b(\Phi(x))}{a(\Phi(x))} + \frac{v}{a(\Phi(x))} \\ &= \frac{-L_f^n h(x) + v}{L_g L_f^{n-1} h(x)} \end{aligned} \tag{17.89}$$

which corresponds to the same controllable linear system (17.87). This control law is called linearizing state feedback and the coordinates Φ are the linearizing coordinates.

Two remarks (Isidori, 1995) are particularly important:

– We assumed that x° is a stationary point for the system (17.2), thus $f(x^\circ) = 0$ and $h(x^\circ) = 0$, hence

$$\begin{aligned} \phi_1(x^\circ) &= h(x^\circ) = 0 \\ \phi_i(x^\circ) &= \frac{\partial L_f^{i-2} h}{\partial x} f(x^\circ) = 0 \qquad , \quad 1 < i \leq n \end{aligned} \tag{17.90}$$

or $z^\circ = 0$. It is always possible to come back to $h(x^\circ) = 0$ by an appropriate translation.

– It is possible to realize a pole-placement or to satisfy an optimality criterion, by imposing a feedback (Fig. 17.5) as

$$v_2 = K\, z \qquad , \quad \text{with the gain vector: } K = [c_0 \ldots c_{n-1}] \tag{17.91}$$

equivalent to

$$v_2 = c_0 h(x) + c_1 L_f h(x) + \ldots + c_{n-1} L_f^{n-1} h(x) \tag{17.92}$$

which is a nonlinear state feedback with respect to x. The state feedback control law becomes, in this case,

$$u(t) = \frac{-L_f^n h(x) + \dot{z}_n}{L_g L_f^{n-1} h(x)} = \frac{-L_f^n h(x) - \sum_{i=0}^{n-1} c_i L_f^i h(x) + v}{L_g L_f^{n-1} h(x)} \tag{17.93}$$

The external input v, for example, could be equal to the set point y_{ref}. When $v = 0$, this corresponds to the local asymptotic equilibrium $z = 0$ which is preserved.

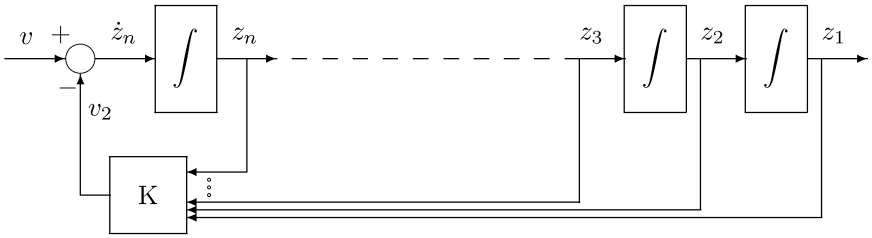


Figure 17.5: Nonlinear control with pole-placement for a system of relative degree equal to n

By explaining the transfer function $Y(s)/V(s)$, the characteristic polynomial is then equal to

$$c_0 + c_1\, s + \ldots + c_{n-1}\, s^{n-1} + s^n \tag{17.94}$$

whose coefficients can be chosen so as to realize the adequate pole-placement.

By simply considering the system (17.81) without output, Isidori (1995) shows that the exact input-state linearization is possible in a neighbourhood U of x° if and only if there exists a scalar function $\lambda(x)$ such that the system with the "output" redefined

$$\begin{array}{ll} \dot{x} & = f(x) + g(x)\, u \\ y & = \lambda(x) \end{array} \tag{17.95}$$

has a relative degree equal to n at x°. The function $\lambda(x)$ is equal to $z_1(x)$.

It is possible to obtain the following general theorem (Hunt et al., 1983; Su, 1982):

Theorem:

the system (17.81) is exactly linearizable in state space (input-state linearization) in a neighbourhood U of x° if and only if the following conditions are satisfied:

1. The vector fields $\{g(x^\circ), ad_f g(x^\circ), \ldots, ad_f^{n-1} g(x^\circ)\}$ are linearly independent.
2. The distribution $\text{span}\{g, ad_f g, \ldots, ad_f^{n-2} g\}$ is involutive in U.

Condition 1 that can be written as "the following matrix

$$[g(x^\circ), ad_f g(x^\circ), \ldots, ad_f^{n-1} g(x^\circ)]$$

has rank n" is a controllability condition of the nonlinear system. This matrix must be invertible. In the linear context, this matrix is the controllability matrix

$$[B, AB, \ldots, A^{n-1}B]$$

To realize the exact input-state linearization, we must proceed according to the following stages:

- Build the vector fields $g(x^\circ), ad_f g(x^\circ), \ldots, ad_f^{n-1} g(x^\circ)$.
- Check if the conditions of controllability and involutivity are verified.
- If these conditions are verified, find the function $\lambda(x)$ from the equations
$$L_g \lambda(x^\circ) = L_g L_f \lambda(x^\circ) = \ldots = L_g L_f^{n-2} \lambda(x^\circ) = 0$$
$$L_g L_f^{n-1} \lambda(x^\circ) \neq 0$$
- Calculate the coordinate change

$$\Phi(x) = [\lambda(x), L_f \lambda(x), \ldots, L_f^{n-1} \lambda(x)] \tag{17.96}$$

17.2.9 Input-Output Linearization of a System with Relative Degree r Lower than or Equal to n

Two cases are distinguished:

- (Isidori, 1995) notices that a nonlinear system such as

$$\left\{ \begin{array}{ll} \dot{x} & = f(x) + g(x)\, u \\ y & = h(x) \end{array} \right. \tag{17.97}$$

having a relative degree $r < n$ can satisfy the conditions of the previous theorem. In this case, there exists a different "output" λ such that the system has a relative degree equal to n. The new system thus defined satisfies the previous theorem; by using a feedback $u = \alpha(x) + \beta(x)\, v$ and a coordinate change $\Phi(x)$, it is transformed into a controllable linear system, but the real output, in general, is not linear with respect to the new

$$y = h(\Phi^{-1}(z)) \tag{17.98}$$

- If the output y is fixed by $y = h(x)$ and if the system possesses a relative degree r lower than or equal to n, by using the coordinate change $\Phi(x)$, it is possible to transform the system into the normal form (17.77) and to set $v = \dot{z}_r$ so that the system is simply expressed in the transformed coordinates in Byrnes-Isidori canonical form

$$\begin{array}{ll} \dot{z}_1 & = z_2 \\ & \vdots \\ \dot{z}_{r-1} & = z_r \\ \dot{z}_r & = v = b(z) + a(z)\, u(t) \\ \dot{z}_{r+1} & = q_{r+1}(z) \\ \dot{z}_n & = q_n(z) \\ y = z_1 & \end{array} \tag{17.99}$$

with

$$b(z) = L_f^r h(x) \quad , \quad a(z) = L_g L_f^{r-1} h(x) \tag{17.100}$$

The control law is deduced

$$\begin{aligned} u(t) &= -\frac{b(z)}{a(z)} + \frac{v}{a(z)} \\ &= \frac{-L_f^r h(x) + v}{L_g L_f^{r-1} h(x)} \end{aligned} \tag{17.101}$$

The resulting system is only partially linear, but the output is influenced by the external input v only through a chain of r integrators (Fig. 17.2) related to the new states $z_1, \ldots, z_r$

$$y^{(r)} = L_f^r h(x) + L_g L_f^{r-1} h(x)\, u = v \tag{17.102}$$

The new states $z_{r+1}, \ldots, z_n$ which constitute the nonlinear part of the system do not influence the output y.

17.2.10 Zero Dynamics

In the case of a linear system of relative degree $r = n$, the transfer function does not possess any zeros. Similarly, for a nonlinear system, for that purpose, we consider only the case where the relative degree r is lower than n. We represent

the vector in normal form (17.77) by separating the linear part of dimension r and the nonlinear part of dimension $n - r$, thus

$$\xi = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_r \end{bmatrix} = \begin{bmatrix} y \\ \dot{y} \\ \vdots \\ y^{(r-1)} \end{bmatrix} \quad ; \quad \eta = \begin{bmatrix} z_{r+1} \\ \vdots \\ z_n \end{bmatrix} \tag{17.103}$$

allowing us to rewrite the system as

$$\begin{array}{ll} \dot{z}_1 & = z_2 \\ & \vdots \\ \dot{z}_{r-1} & = z_r \\ \dot{z}_r & = b(\xi, \eta) + a(\xi, \eta)\, u(t) \\ \dot{\eta} & = q(\xi, \eta) \end{array} \tag{17.104}$$

where ξ and η constitute the normal coordinates or normal states.

The dynamics of the nonlinear system is thus decomposed into an external input-output part and an internal unobservable part. It is simple to conceive the external part, but there remains the problem of the internal stability corresponding to the last $(n - r)$ equations: $\dot{\eta} = q(\xi, \eta)$.

As x° is an equilibrium point of the system, we obtain $f(x^\circ) = 0$ and we can choose $h(x^\circ) = 0$. We can assume that, in the normal coordinates, the point $(0, 0)$ is the equilibrium point, hence $b(0, 0) = 0$ and $q(0, 0) = 0$.

We seek to make the output zero for all t in the neighbourhood of $t = 0$. In the normal form, this would amount to imposing

$$\dot{z}_1 = \ldots = \dot{z}_r = 0 \Longleftrightarrow \xi = 0 \qquad \text{for all } t \tag{17.105}$$

as, moreover, we maintain $y = z_1 = 0$. The input u results such that

$$0 = b(0, \eta) + a(0, \eta)\, u(t) \tag{17.106}$$

with $a(0, \eta) \neq 0$ still in the neighbourhood of $t = 0$. Moreover, the variable η is such that

$$\dot{\eta} = q(0, \eta) \qquad , \quad \text{with: } \eta(0) = \eta^\circ \tag{17.107}$$

which is an autonomous system of differential equations whose solution is the variable $\eta(t)$. We deduce the unique expression of the input that imposes a zero output in the neighbourhood of $t = 0$

$$u(t) = -\frac{b(0, \eta(t))}{a(0, \eta(t))} \tag{17.108}$$

The dynamics of Eq. (17.107), which results from the condition of zero output, is called zero dynamics or unforced zero dynamics; it describes the internal behaviour of the system.

The search of the zero output could have been realized in the original state space, by setting

$$y(t) = \dot{y}(t) = \ldots = y^{(r-1)}(t) = y^{(r)}(t) = 0 \qquad \text{for all } t \tag{17.109}$$

or further, in the neighbourhood of x°

$$\begin{array}{l} h(x) = L_f h(x) = \ldots = L_f^{r-1} h(x) = 0 \\ L_f^r h(x) + L_g L_f^{r-1} h(x)\, u(t) = 0 \end{array} \tag{17.110}$$

The case of tracking a reference output y_{ref} is deducted from the previous case of a zero output. We set in the neighbourhood of $t = 0$

$$y(t) = y_{\text{ref}}(t) \tag{17.111}$$

which gives, for the new coordinates

$$z_i(t) = y_{\text{ref}}^{(i-1)}(t) \qquad , \quad 1 \leq i \leq r \tag{17.112}$$

By analogy with the previous case, we set

$$\xi_{\text{ref}} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_r \end{bmatrix} = \begin{bmatrix} y_{\text{ref}} \\ y_{\text{ref}}^{(1)} \\ \vdots \\ y_{\text{ref}}^{(r-1)} \end{bmatrix} \tag{17.113}$$

The equation that imposes the control results

$$y_{\text{ref}}^{(r)}(t) = b(\xi_{\text{ref}}(t), \eta(t)) + a(\xi_{\text{ref}}(t), \eta(t))\, u(t) \tag{17.114}$$

where η is the solution of the autonomous differential system

$$\dot{\eta}(t) = q(\xi_{\text{ref}}(t), \eta(t)) \qquad , \quad \text{with: } \eta(0) = \eta^\circ \tag{17.115}$$

We draw the equation of the unique control imposing on the output to exactly follow the reference

$$u(t) = \frac{y_{\text{ref}}^{(r)}(t) - b(\xi_{\text{ref}}(t), \eta(t))}{a(\xi_{\text{ref}}(t), \eta(t))} \tag{17.116}$$

The system of differential Eqs. (17.115) coupled with Eq. (17.116) gives the forced zero dynamics or dynamics of the inverse of the system (17.97), corresponding to a control such that the output exactly follows the reference. η is the state of the dynamics of the inverse, ξ_{ref} its control input and u its output.

17.2.11 Asymptotic Stability

We consider the system in its normal form

$$\begin{array}{ll} \dot{z}_1 & = z_2 \\ & \vdots \\ \dot{z}_{r-1} & = z_r \\ \dot{z}_r & = b(\xi,\eta) + a(\xi,\eta)\, u(t) \\ \dot{\eta} & = q(\xi,\eta) \end{array} \tag{17.117}$$

by considering as previously that $(\xi, \eta) = (0, 0)$ is an equilibrium point. We consider a state feedback close to Eq. (17.93), in this case

$$v = -Kz \qquad , \quad \text{with the gain vector: } K = [c_0 \dots c_{r-1}] \tag{17.118}$$

The state feedback

$$\begin{array}{rl} u(t) & = \dfrac{-b(\xi,\eta) - \sum_{i=0}^{r-1} c_i z_{i+1}}{a(\xi,\eta)} \\ & = \dfrac{-L_f^r h(x) - \sum_{i=0}^{r-1} c_i L_f^i h(x)}{L_g L_f^{r-1} h(x)} \end{array} \tag{17.119}$$

gives the closed-loop system

$$\begin{array}{ll} \dot{\xi} & = A\xi \\ \dot{\eta} & = q(\xi,\eta) \end{array} \tag{17.120}$$

where A is equal to

$$\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & 1 & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \dots & & 0 & 1 \\ -c_0 & -c_1 & \dots & \dots & -c_{r-1} \end{bmatrix} \tag{17.121}$$

which is a companion controllability matrix and has the characteristic polynomial

$$c_0 + c_1\, s + \ldots + c_{r-1}\, s^{r-1} + s^r \tag{17.122}$$

If on one hand, the coefficients are chosen so that the roots of this polynomial have a negative real part and, on the other hand, the zero dynamics corresponding to $\dot{\eta} = q(0, \eta)$ is asymptotically locally stable, the state feedback (17.119) stabilizes asymptotically locally the system (17.120) in the neighbourhood of the equilibrium $(\xi, \eta) = (0, 0)$.

The role and the importance of zero dynamics thus appear clearly here. If the linear approximation of the system possesses uncontrollable modes, the

latter necessarily correspond to eigenvalues of the linear approximation Q of the zero dynamics. The linear approximation of the system is given by

$$\begin{array}{ll} \dot{z}_1 & = z_2 \\ & \vdots \\ \dot{z}_{r-1} & = z_r \\ \dot{z}_r & = R\,\xi + S\,\eta + K\,u \\ \dot{\eta} & = P\,\xi + Q\,\eta \end{array} \tag{17.123}$$

with the partial derivative matrices considered at $(\xi, \eta) = (0, 0)$

$$R = \left[\frac{\partial b}{\partial \xi}\right] \quad , \quad S = \left[\frac{\partial b}{\partial \eta}\right] \quad , \quad P = \left[\frac{\partial q}{\partial \xi}\right] \quad , \quad Q = \left[\frac{\partial q}{\partial \eta}\right] \tag{17.124}$$

Notice that it is not necessary that the linear approximation is asymptotically stable for the nonlinear system to be stable.

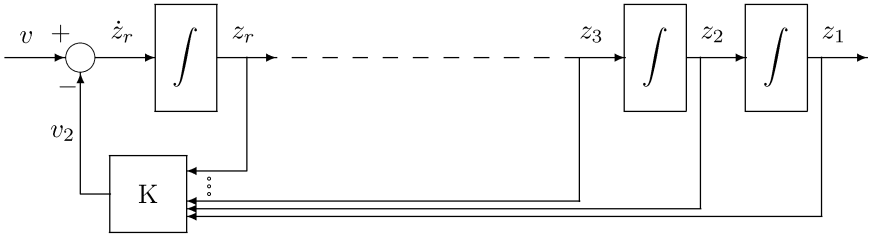


Figure 17.6: Nonlinear control with pole-placement for a system of relative degree $r \leq n$

As has already been realized for a system of relative degree n with the state feedback (17.93), we can take into account an external input v (Fig. 17.6) as

$$u(t) = \frac{-L_f^r h(x) - \sum_{i=0}^{r-1} c_i L_f^i h(x) + v}{L_g L_f^{r-1} h(x)} \tag{17.125}$$

so that the system (17.117) is transformed into

$$\begin{array}{ll} \dot{\xi} & = A\xi + B\,v \\ \dot{\eta} & = q(\xi, \eta) \end{array} \tag{17.126}$$

with: $B = [0 \ldots 01]^T$. Provided that the zero dynamics is stable, the stability will depend on the characteristic polynomial

$$c_0 + c_1\,s + \ldots + c_{r-1}\,s^{r-1} + s^r \tag{17.127}$$

whose coefficients will be chosen so as to realize the desired pole-placement.

By analogy with the linear systems, a nonlinear system is called minimum-phase if its unforced zero dynamics is asymptotically locally stable at $(0, 0)$.

17.2.12 Tracking of a Reference Trajectory

So that the output $y = z_1$ converges asymptotically towards a reference trajectory y_{ref}, it is suitable to gather the elements of both previous sections: the forced zero dynamics with reference trajectory and the asymptotic stability.

We still consider the system in its normal form (17.117), and the state feedback is

$$\begin{aligned} u(t) &= \frac{-b(\xi,\eta) + y_{\text{ref}}^{(r)} - \sum_{i=0}^{r-1} c_i (z_{i+1} - y_{\text{ref}}^{(i)})}{a(\xi,\eta)} \\ &= \frac{-L_f^r h(x) + y_{\text{ref}}^{(r)} - \sum_{i=0}^{r-1} c_i (L_f^i h(x) - y_{\text{ref}}^{(i)})}{L_g L_f^{r-1} h(x)} \end{aligned} \tag{17.128}$$

The error defined by

$$e(t) = y(t) - y_{\text{ref}}(t) \tag{17.129}$$

is the solution of the following differential equation

$$e^{(r)} + c_{r-1}\, e^{(r)} + \ldots + c_1\, e^{(1)} + c_0\, e = 0 \tag{17.130}$$

whose parallel with the characteristic polynomial (17.127) is obvious. By choosing in an adequate manner the coefficients c_i, it is possible to ensure the exponential convergence of the error towards 0 when $t \to \infty$. In the same way as the previous section, it is necessary that the zero dynamics (here forced) corresponding to

$$\dot{\eta} = q(\xi_{\text{ref}}, \eta) \qquad , \quad \text{with: } \eta_{\text{ref}}(0) = 0 \tag{17.131}$$

is stable (where $\eta_{\text{ref}}(t)$ is the solution of this differential system) and with the vector

$$\xi_{\text{ref}} = [y_{\text{ref}}(t), y_{\text{ref}}^{(1)}(t), \ldots, y_{\text{ref}}^{(r-1)}(t)]^T \tag{17.132}$$

Isidori (1995) studied the particular case where the reference y_{ref} depends on a linear model. In the case where an external input is taken into account, the state feedback (17.128) becomes

$$u(t) = \frac{v - L_f^r h(x) + y_{\text{ref}}^{(r)} - \sum_{i=0}^{r-1} c_i (L_f^i h(x) - y_{\text{ref}}^{(i)})}{L_g L_f^{r-1} h(x)} \tag{17.133}$$

17.2.13 Decoupling with Respect to a Disturbance

We assume that a modelable disturbance d acts on the system thus reformulated

$$\left\{ \begin{array}{ll} \dot{x} & = f(x) + g(x)\, u + w(x)\, d \\ y & = h(x) \end{array} \right. \tag{17.134}$$

We wish to define an input u by static state feedback such that the output y does not depend (is decoupled) on the disturbance d. Express the system with the normal coordinates

$$\dot{z}_1 = L_f h(x(t)) + L_g h(x(t))\, u(t) + L_w h(x(t))\, d(t) \tag{17.135}$$

We use, on one hand, the relative degree r of the system, which gives $L_g h = 0$, if $r > 1$, and we impose $L_w h = 0$ so that the disturbance has no influence. For the following coordinates, a similar condition

$$L_w L_f^{i-1} h = 0 \qquad , \quad 1 \leq i \leq r \tag{17.136}$$

is used, hence the system in the normal form including the influence of the disturbance from rank $r+1$

$$\begin{array}{ll} \dot{z}_1 & = z_2 \\ & \vdots \\ \dot{z}_{r-1} & = z_r \\ \dot{z}_r & = L_f^r h(x(t)) + L_g L_f^{r-1} h(x(t))\, u(t) = b(\xi,\eta) + a(\xi,\eta)\, u \\ \dot{\eta} & = q(\xi,\eta) + k(\xi,\eta)\, d \end{array} \tag{17.137}$$

with, moreover $y = z_1$. It clearly appears from the normal form (17.137) that the input defined by the static state feedback

$$u = \frac{-b(\xi,\eta) + v}{a(\xi,\eta)} \tag{17.138}$$

which gives

$$\dot{z}_r = v \tag{17.139}$$

perfectly decouples the output $y = z_1$ from the disturbance d.

The decoupling condition is thus simply Eq. (17.136), which can be expressed as

$$\begin{array}{l} < L_f^{i-1} h, w >= 0 \qquad , \quad 1 \leq i \leq r \Longleftrightarrow \\ w(x) \text{ belongs to the codistribution } \Delta^{\perp} \text{ in the neighbourhood of } x^{\circ} \end{array} \tag{17.140}$$

where the distribution Δ is equal to

$$\Delta = \text{span}\{Dh, DL_f h, \ldots, DL_f^{r-1} h\} \tag{17.141}$$

The previous decoupling has been realized by state feedback. It is well known that when the disturbance can be measured it is advantageous to consider a feedforward term in the control law, such as

$$u(t) = \alpha(x) + \beta(x)\, v + \gamma(x)\, d \tag{17.142}$$

hence the closed-loop system equations

$$\begin{array}{ll} \dot{x} & = f(x) + g(x)[\alpha(x) + \beta(x)\, v] + [w(x) + g(x)\gamma(x)]\, d \\ y & = h(x) \end{array} \tag{17.143}$$

We can think in a way quite parallel to the decoupling by state feedback. It suffices that

$$\begin{array}{l} < L_f^{i-1} h, w + g\gamma >= 0 \qquad , \quad 1 \leq i \leq r \qquad \Longleftrightarrow \\ [w(x) + g(x)\gamma(x)] \text{ belongs to the codistribution } \Delta^{\perp} \text{ in the neighbourhood of } x^{\circ} \end{array} \tag{17.144}$$

This condition can be simplified by explaining the Lie bracket

$$\begin{aligned} < L_f^{i-1}h, w + g\gamma > &= L_{w+g\gamma} L_f^{i-1} h(x) \\ &= L_w L_f^{i-1} h(x) + L_g L_f^{i-1} h(x)\gamma(x) = 0 \quad , \quad 1 \leq i \leq r \end{aligned} \tag{17.145}$$

We draw the value of the function γ

$$\gamma(x) = -\frac{L_w L_f^{r-1} h(x)}{L_g L_f^{r-1} h(x)} \tag{17.146}$$

which is to be reinjected in control law (17.142).

17.2.14 Case of Nonminimum-Phase Systems

Rigorously, the nonlinear control law (17.128) cannot be applied to nonminimum-phase systems, as they are not invertible. Nevertheless, it is possible (Slotine and Li, 1991) to apply to them a control law that gives a small tracking error or to redefine the output so that the modified zero dynamics is stable.

17.2.15 Globally Linearizing Control

This linearizing control of an input-output type proposed (Kravaris, 1988; Kravaris and Chung, 1987; Kravaris and Kantor, 1990a,b; Kravaris and Soroush, 1990) relies on the same concepts of differential geometry as those which have been developed in the previous sections and can be considered as a close variant. The formulation of the control law applied to a minimum-phase system of relative degree r is very close to Eq. (17.128) and is equal to

$$u = \frac{v - L_f^r h(x) - \beta_1 L_f^{r-1} h(x) - \ldots - \beta_{r-1} L_f h(x) - \beta_r h(x)}{L_g L_f^{r-1} h(x)} \tag{17.147}$$

which gives the following input-output linear dynamics

$$y^{(r)} + \beta_1 \, y^{(r-1)} + \ldots + \beta_{r-1} \, y^{(1)} + \beta_r \, y = v \tag{17.148}$$

In order to guarantee a zero asymptotic error, even in the presence of modelling errors and step disturbances (for a PI), the external input can be provided by the following controller

$$v(t) = \int_0^t c(t-\tau) \, \left[y_{\text{ref}}(\tau) - y(\tau) \right] \, d\tau \tag{17.149}$$

where the function $c(t)$, for example, can be chosen as the inverse of a given transfer function. Frequently, v will be a PI controller (Fig. 17.7), e.g.

$$v(t) = K_c \left[y_{\text{ref}}(t) - y(t) + \frac{1}{\tau_i} \int_0^t \left(y_{\text{ref}}(\tau) - y(\tau) \right) d\tau \right] \tag{17.150}$$

In this case, the system stability is conditioned by the roots of the characteristic polynomial

$$s^{r+1} + \beta_1 s^r + \ldots + \beta_{r-1} s^2 + (\beta_r + K_c) s + \frac{K_c}{\tau_I} = 0 \tag{17.151}$$

This control has been tested in different forms such as: combination with a feedforward controller (Daoutidis and Kravaris, 1989) to take into account disturbances, usage of a Smith predictor and a state observer to take into account time delays (Kravaris and Wright, 1989), modification to be applied to nonminimum-phase second-order systems (Kravaris and Daoutidis, 1990), application to a polymerization reactor (Soroush and Kravaris, 1992), with a robustness study (Kravaris and Palanki, 1988).

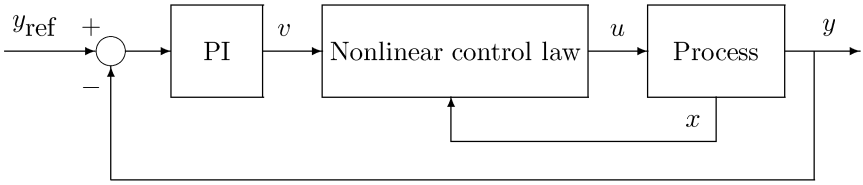


Figure 17.7: Globally Linearizing Control with PI controller

Generic model control (Lee, 1993) proposed by Lee and Sullivan (1988) has not been formulated by using the concepts of differential geometry. However, it can be considered in this framework. Indeed, it is based on a realization of the inverse of the model. For this reason, its applicability is limited (Henson and Seborg, 1990a), as it is reserved to systems of relative degree equal to 1.

Certain authors (Bequette, 1991; Henson and Seborg, 1989, 1990b) compare globally linearizing control (GLC) and generic model control (GMC) to nonlinear control developed by differential geometry, as has been discussed by Isidori (1995). We can consider that GLC and GMC represent minor variants with respect to the general scope, which has been resumed in the previous sections.

17.3 Multivariable Nonlinear Control

Geometric nonlinear control can be extended to multi-input multi-output systems; it will be presented only for systems with the same number of inputs and outputs. The details of the demonstrations can be found, in particular, in the textbook by (Isidori, 1995). Only the main points will be cited here.

Multivariable nonlinear systems, affine with respect to the inputs, are modelled as

$$\begin{array}{ll} \dot{x} & = f(x) + \sum_{i=1}^{m} g_i(x)\, u_i \\ y_i & = h_i(x) \qquad , \quad i = 1, \ldots, m \end{array} \tag{17.152}$$

with

$$u = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \quad ; \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \tag{17.153}$$

$f(x), g_i(x), h_i(x)$ are smooth vector fields. In a more compact form, the system is modelled by

$$\begin{aligned} \dot{x} &= f(x) + g(x)\, u \\ y &= h(x) \end{aligned} \tag{17.154}$$

17.3.1 Relative Degree

Many reasonings used to define the concepts for multivariable systems are an extension of the single-input single-output case. Some of them are specific. In the multivariable case, the notion of relative degree is extended by

– The vector of relative degrees $[r_1, \ldots r_m]^T$ defined in a neighbourhood of x° by

$$L_{g_j} L_f^k h_i(x) = 0 \quad , \quad \forall j = 1, \ldots, m \quad , \quad \forall k < r_i - 1 \quad , \quad \forall i = 1, \ldots, m \tag{17.155}$$

– The matrix $A(x)$ defined by

$$A(x) = \begin{bmatrix} L_{g_1} L_f^{r_1 - 1} h_1(x) & \ldots & L_{g_m} L_f^{r_1 - 1} h_1(x) \\ \vdots & & \vdots \\ L_{g_1} L_f^{r_m - 1} h_m(x) & \ldots & L_{g_m} L_f^{r_m - 1} h_m(x) \end{bmatrix} \tag{17.156}$$

which must be nonsingular at x°.

If we consider a given output of subscript i, we observe that the vector of the Lie derivatives verifies

$$\begin{aligned} &\begin{bmatrix} L_{g_1} L_f^k h_i(x) & \ldots & L_{g_m} L_f^k h_i(x) \end{bmatrix} = 0 \quad \forall k < r_i - 1 \\ &\begin{bmatrix} L_{g_1} L_f^{r_i - 1} h_i(x) & \ldots & L_{g_m} L_f^{r_i - 1} h_i(x) \end{bmatrix} \neq 0 \end{aligned} \tag{17.157}$$

where r_i is the number of times that the output $y_i(t)$ must be differentiated to make at least one component of the vector $u(t)$ appear. There exists at least one couple (u_j, y_i) that have r_i as the relative degree.

Now consider a system having $[r_1, \ldots r_m]^T$ as the vector of relative degrees. The vectors

$$\begin{aligned} &\begin{bmatrix} Dh_1(x^\circ), & DL_f h_1(x^\circ), & \ldots, & DL_f^{r_1 - 1} h_1(x^\circ) \end{bmatrix} \\ &\vdots \\ &\begin{bmatrix} Dh_m(x^\circ), & DL_f h_m(x^\circ), & \ldots, & DL_f^{r_m - 1} h_m(x^\circ) \end{bmatrix} \end{aligned} \tag{17.158}$$

are linearly independent.

17.3.2 Coordinate Change

The proposed coordinate change is quite parallel to the single-input single-output case. Consider a system of relative degree vector $[r_1, \ldots r_m]^T$. Let the coordinate change

$$\forall\, 1 \leq i \leq m \qquad \begin{array}{ll} \phi_1^i(x) & = h_i(x) \\ \phi_2^i(x) & = L_f h_i(x) \\ & \vdots \\ \phi_{r_i}^i(x) & = L_f^{r_i - 1} h_i(x) \end{array} \tag{17.159}$$

Denote by $r = r_1 + \ldots + r_m$ the total relative degree.

If $r < n$, it is possible to find $(n - r)$ additional functions $\phi_{r+1}(x), \ldots, \phi_n(x)$ such that the mapping

$$\Phi(x) = \left[\phi_1^1(x), \ldots, \phi_{r_1}^1(x), \ldots, \phi_1^m(x), \ldots, \phi_{r_m}^m(x), \phi_{r+1}(x), \ldots, \phi_n(x)\right]^T \tag{17.160}$$

has its nonsingular Jacobian matrix at x° and thus constitutes a coordinate change in a neighbourhood U of x°.

Moreover, if the distribution

$$G = \text{span}\,\{g_1, \ldots, g_m\} \tag{17.161}$$

is involutive in U, the additional functions: $\phi_{r+1}(x), \ldots, \phi_m(x)$ can be chosen such that

$$L_{g_j} \phi_i(x) = 0 \quad , \quad r + 1 \leq i \leq n \quad , \quad 1 \leq j \leq m \tag{17.162}$$

The condition of nonsingularity of matrix $A(x)$ can be extended to a system having more inputs than outputs; it becomes a rank condition: the matrix rank must be equal to the number of its rows, or, further, the system must have more inputs than outputs, which is classical.

17.3.3 Normal Form

The coordinate change gives for ϕ_i^1 (similarly for the other ϕ_i^j)

$$\begin{array}{ll} \dot{\phi}_1^1 & = \phi_2^1(x) \\ & \vdots \\ \dot{\phi}_{r_1-1}^1 & = \phi_{r_1}^1(x) \\ \dot{\phi}_{r_1-1}^1 & = L_f^{r_1} h_1(x) + \sum_{j=1}^m L_{g_j} L_f^{r_1 - 1} h_1(x)\, u_j(t) \end{array} \tag{17.163}$$

The normal coordinates are introduced

$$\xi = [\xi^1, \ldots, \xi^m] \quad \text{with:} \quad \xi^i = \begin{bmatrix} \xi_1^i \\ \vdots \\ \xi_{r_i}^i \end{bmatrix} = \begin{bmatrix} \phi_1^i(x) \\ \vdots \\ \phi_{r_i}^i(x) \end{bmatrix}, i = 1, \ldots, m$$
$$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_{n-r} \end{bmatrix} = \begin{bmatrix} \phi_{r+1}(x) \\ \vdots \\ \phi_n^i(x) \end{bmatrix} \tag{17.164}$$

which gives the normal form

$$\begin{array}{ll} \dot{\xi}_1^i & = \xi_2^i \\ & \vdots \\ \dot{\xi}_{r_i-1}^i & = \xi_{r_i}^i \\ \dot{\xi}_{r_i}^i & = b_i(\xi, \eta) + \sum_{j=1}^m a_{ij}(\xi, \eta)\, u_j \\ \dot{\eta} & = q(\xi, \eta) + p(\xi, \eta)\, u \\ y & = \xi_1^i \end{array} \tag{17.165}$$

with

$$\begin{array}{ll} a_{ij}(\xi, \eta) & = L_{g_j} L_f^{r_i - 1} h_i(\Phi^{-1}(\xi, \eta)) \quad , \; i, j = 1, \ldots, m \\ b_i(\xi, \eta) & = L_f^{r_i} h_i(\Phi^{-1}(\xi, \eta)) \quad , \; i = 1, \ldots, m \end{array} \tag{17.166}$$

where a_{ij} are the coefficients of matrix $A(x)$.

17.3.4 Zero Dynamics

The zero dynamics is defined in the same way as for single-input single-output systems. The inputs and the initial conditions are sought so that the outputs are identically zero in a neighbourhood U of x°. The control vector must be such that

$$u(t) = -A^{-1}(0, \eta(t))\, b(0, \eta(t)) \tag{17.167}$$

and the zero dynamics or unforced dynamics is defined by the differential system

$$\dot{\eta}(t) = q(\xi, \eta) - p(\xi, \eta)\, A^{-1}(\xi, \eta)\, b(\xi, \eta) \qquad , \quad \eta(0) = \eta^\circ \tag{17.168}$$

The control (17.167) could be expressed in the original state space as

$$u^*(x) = -A^{-1}(x)\, b(x) \tag{17.169}$$

The change from the unforced dynamics to the forced dynamics would be realized in the same manner as for single-input single-output systems.

17.3.5 Exact Linearization by State Feedback and Diffeomorphism

Consider the system

$$\dot{x} = f(x) + g(x)\, u \tag{17.170}$$

without taking into account the outputs.

We introduce the distributions

$$\begin{aligned} G_0 &= \text{span}\,\{g_1, \ldots, g_m\} \\ G_1 &= \text{span}\,\{g_1, \ldots, g_m, \text{ad}_f g_1, \ldots, \text{ad}_f g_m\} \\ &\vdots \\ G_i &= \text{span}\left\{\text{ad}_f^k g_1, \ldots, \text{ad}_f^k g_m \,;\, 0 \le k \le i\right\} \qquad ; \quad i = 0, \ldots, n-1 \end{aligned} \tag{17.171}$$

The matrix $g(x^\circ)$ is assumed of rank m. The exact linearization is possible if and only if:

- The distribution G_i , $(i = 0, \ldots, n-1)$ has a constant dimension in the neighbourhood of x°.
- The distribution G_{n-1} has a dimension equal to n.
- The distribution G_i , $(i = 0, \ldots, n-2)$ is involutive.

Several facts must be looked at:

- The non-singularity of matrix $A(x)$ (Eq. 17.156).
- The total relative degree r must be equal to n.
- As for single-input single-output systems, the outputs y_i are given by the solutions $\lambda_i(x)$, $(j = 1, \ldots, m)$ of the equations

$$L_{g_j} L_f^k \lambda_i(x) = 0 \qquad , \quad 0 \le k \le r_i - 2 \;,\; j = 1, \ldots, m \tag{17.172}$$

The linearizing state feedback is

$$u = -A^{-1}(x)\, b(x) + A^{-1}(x)\, v \tag{17.173}$$

and the linearizing normal coordinates are

$$\xi_k^i(x) = L_f^{k-1} h_i(x) \qquad , \quad 1 \le k \le r_i \;,\; i = 1, \ldots, m \tag{17.174}$$

17.3.6 Nonlinear Control Perfectly Decoupled by Static State Feedback

The decoupling for the system (17.154) will be perfectly realized when any output y_i $(1 \le i \le m)$ is influenced only by the corresponding input v_i. This problem has a solution only if the decoupling matrix $A(x)$ is nonsingular at x°, i.e. if the system possesses a vector of relative degrees.

The static state feedback is equal to

$$u = -A^{-1}(x)\, b(x) + A^{-1}(x)\, v \tag{17.175}$$

To display the decoupling, it suffices to consider the system in its normal form (17.165) and to propose the following control law

$$u = -A^{-1}(\xi,\eta)\, b(\xi,\eta) + A^{-1}(\xi,\eta)\, v \tag{17.176}$$

which transforms the system into

$$\begin{array}{ll}
\dot{\xi}_1^i & = \xi_2^i \\
& \vdots \\
\dot{\xi}_{r_i-1}^i & = \xi_{r_i}^i \\
\dot{\xi}_{r_i}^i & = b_i(\xi,\eta) + \sum_{j=1}^m a_{ij}(\xi,\eta)\, u_j \\
& = L_f^{r_i} h_i(\Phi^{-1}(\xi,\eta)) + \sum_{j=1}^m L_{g_j} L_f^{r_i-1} h_i(\Phi^{-1}(\xi,\eta))\, u_j \\
& = v_i \\
\dot{\eta} & = q(\xi,\eta) + p(\xi,\eta)\, u \\
y & = \xi_1^i
\end{array} \tag{17.177}$$

as we could write

$$\begin{bmatrix} \dot{\xi}_{r_1}^1 \\ \cdots \\ \dot{\xi}_{r_m}^m \end{bmatrix} = b(\xi,\eta) + A(\xi,\eta)\, u = v \tag{17.178}$$

from Eq. (17.166), giving the coefficients a_{ij}, the control law (17.176) and the normal form.

Two cases can occur:

- The total relative degree r is lower than n; there then exists an unobservable part in the system, influenced by the inputs and the states, but not influencing the outputs.
- The total relative degree r is equal to n; the system can be decoupled into m chains, each composed of r_i integrators. The system is thus transformed into a completely linear and controllable system.

Of course, we must verify that the decoupling matrix $A(x)$ is nonsingular at x°.

Just as for single-input single-output systems, it is possible to realize a pole-placement by adding a state feedback as

$$v_i = -c_0^i \xi_1^i - \ldots - c_{r_i-1}^i \xi_{r_i}^i \tag{17.179}$$

The different points treated for single-input single-output systems, asymptotic stability of the zero dynamics, disturbance rejection, reference model tracking, are studied in a very similar manner.

Example 17.1: **Multivariable Nonlinear Control of an Evaporator**

(To et al., 1995) describe nonlinear control of a simulated industrial evaporator by means of different techniques, including input-output linearization, Su-Hunt-Meyer transformation and generic model control. The model of the process has three states, two manipulated inputs and two controlled outputs.

17.3.7 Obtaining a Relative Degree by Dynamic Extension

Some systems possess no relative degree vector. No static state feedback can change this result, as the relative degree property is invariant by this means.

Suppose that the system

$$\begin{aligned} \dot{x} &= f(x) + g(x)\, u \\ y &= h(x) \end{aligned} \tag{17.180}$$

possesses no total relative degree, as the rank of matrix $A(x)$ is smaller than m. To obtain the relative degree, a dynamic part is added between the old inputs u and the new inputs v, modelled by

$$\begin{aligned} u &= \alpha(x, \zeta) + \beta(x, \zeta)\, v \\ \dot{\zeta} &= \gamma(x, \zeta) + \delta(x, \zeta)\, v \end{aligned} \tag{17.181}$$

A simple type of dynamics used is the interposition of integrators between an input v_i and the input u_i (Fig. 17.8) modelled in the case of two integrators by

$$\begin{aligned} u_i &= \zeta_1 \\ \dot{\zeta}_2 &= v_i \\ \dot{\zeta}_1 &= \zeta_2 \end{aligned} \tag{17.182}$$

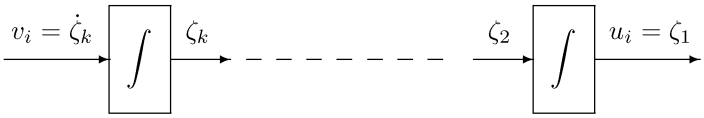


Figure 17.8: Dynamic extension

The modified system will be

$$\begin{aligned} \dot{x} &= f(x) + g(x)\, \alpha(x, \zeta) + g(x)\, \beta(x, \zeta)\, v \\ \dot{\zeta} &= \gamma(x, \zeta) + \delta(x, \zeta)\, v \\ y &= h(x) \end{aligned} \tag{17.183}$$

The algorithm of dynamic extension described by Isidori (1995) consists of progressively increasing the rank of matrix $A(x, \zeta)$ corresponding to the modified system until it is equal to m. In particular, this algorithm contains a procedure of identification of the inputs on which it is necessary to act.

In summary, if the relative degree is obtained by dynamic extension and if the total relative degree of the extended system is equal to n, it is possible to assert that the original system can be transformed into a completely linear and controllable system by a dynamic state feedback and a coordinate change.

Example 17.2: Nonlinear Control of a Continuous Polymerization Reactor with Dynamic Extension

Soroush and Kravaris (1994) describe a continuous polymerization reactor where

conversion and temperature are controlled by manipulating two coordinated flow rates and two coordinated heat inputs. Modeled in this way, the system exhibits a singular characteristic matrix. They used a dynamic input-output linearizing state feedback by redefining the second manipulated input using just the rate of change (thus the derivative) of monomer flow rate instead of the original monomer flow rate. This amounts to adding an integrator.

17.3.8 Nonlinear Adaptive Control

Frequently, the process model is uncertain or time-varying. In this case, we can think of realizing on-line identification of physical parameters influencing the model (Kosanovich et al., 1995). Thus, Wang et al. (1995) studied a batch styrene polymerization reactor. At the beginning of the reaction, the polymerization has made little progress and the viscosity of the reactor contents is near that of the solvent. When the monomer conversion increases, the viscosity strongly increases because of the gel effect, and the heat transfer coefficient decreases significantly because the stirring progressively passes from a turbulent regime to a laminar one, creating reactor runaway hazard. The model used by (Wang et al., 1995) describes these phenomena, which have been taken into account by using an augmented state vector. Besides the traditional states, which are the concentrations and temperatures, the gel effect coefficient and the heat transfer coefficient are estimated during the reaction. The used technique is inspired from procedures of recursive identification presented by Sastry and Bodson (1989). Presently, this research domain is very important Krstić et al. (1995).

17.4 Applications of Nonlinear Geometric Control

Example 17.3: Nonlinear Geometric Control of Chemical and Biochemical Reactors

Two single-input single-output applications of nonlinear geometric control performed in simulation for realistic models are discussed in Chap. 19. The first one concerns a continuous stirred tank reactor where a chemical reaction takes place. The second one concerns a fed-batch biological reactor. These reactors present different relative degrees. For each of them, an extended Kalman filter is used to estimate the states. From these examples, it should easily become apparent to the reader how to apply nonlinear geometric control in a real example, at least for SISO systems. An example of pilot application to a copolymerization reactor preceded by a study of dynamic optimization in order to obtain the optimal profile temperature is given by Gentric et al. (1999).

References

B.W. Bequette. Nonlinear control of chemical processes: a review. *Ind. Eng. Chem. Res.*, 30:1391–1413, 1991.

P. Daoutidis and C. Kravaris. Synthesis of feedforward/state feedback controllers for nonlinear processes. *AIChE J.*, 35(10):1602–1616, 1989.

A.J. Fossard and D. Normand-Cyrot, editors. *Systèmes non Linéaires 3. Commande.* Masson, Paris, 1993.

C. Gentric, F. Pla, M.A. Latifi, and J.P. Corriou. Optimization and non-linear control of a batch emulsion polymerization reactor. *Chem. Eng. J.*, 75:31–46, 1999.

M.A. Henson and D.E. Seborg. A unified differential geometric approach to nonlinear process control. San Francisco, 1989. AIChE Annual Meeting.

M.A. Henson and D.E. Seborg. A critique of differential geometric control strategies for process control. USSR, 1990a. 11th IFAC World Congress.

M.A. Henson and D.E. Seborg. Input-output linearization of general nonlinear processes. 36(11):1753–1757, 1990b.

R.M. Hirschorn. Invertibility of nonlinear control systems. *SIAM. J. Control Optim.*, 17:289–295, 1979.

L.R. Hunt, R. Su, and G. Meyer. Global transformations of nonlinear systems. *IEEE Trans. Automat. Control*, 28:24–31, 1983.

A. Isidori. *Nonlinear Control Systems: An Introduction.* Springer-Verlag, New York, 2nd edition, 1989.

A. Isidori. *Nonlinear Control Systems.* Springer-Verlag, New York, 3rd edition, 1995.

H.K. Khalil. *Nonlinear Systems.* Prentice Hall, 1996.

K.A. Kosanovich, M.J. Piovosa, V. Rokhlenko, and A. Guez. Nonlinear adaptive control with parameter estimation of a CSTR. *J. Proc. Cont.*, 5(3): 137–148, 1995.

C. Kravaris. Input-output linearization : a nonlinear analog of placing poles at process zeros. *AIChE J.*, 34(11):1803–1812, 1988.

C. Kravaris and C.B. Chung. Nonlinear state feedback synthesis by global input/output linearization. *AIChE J.*, 33(4):592–603, 1987.

C. Kravaris and P. Daoutidis. Nonlinear state feedback control of second-order non-minimum phase nonlinear systems. *Comp. Chem. Eng.*, 14:439–449, 1990.

C. Kravaris and J.C. Kantor. Geometric methods for nonlinear process control. 1. background. *Ind. Eng. Chem. Res.*, 29:2295–2310, 1990a.

C. Kravaris and J.C. Kantor. Geometric methods for nonlinear process control. 2. controller synthesis. *Ind. Eng. Chem. Res.*, 29:2310–2323, 1990b.

C. Kravaris and S. Palanki. Robust nonlinear state feedback under structured uncertainty. *AIChE J.*, 34(7):1119–1127, 1988.

C. Kravaris and M. Soroush. Synthesis of multivariable nonlinear controllers by input/output linearization. *AIChE J.*, 36(2):249–264, 1990.

C. Kravaris and R.A. Wright. Deadtime compensation for nonlinear processes. *AIChE J.*, 35(9):1535–1542, 1989.

M. Krstić, I. Kanellakopoulos, and P. Kokotovic. *Nonlinear and Adaptive Control Design*. Wiley Interscience, New York, 1995.

P.L. Lee, editor. *Nonlinear Process Control: Applications of Generic Model Control*. Springer-Verlag, New York, 1993.

P.L. Lee and G.R. Sullivan. Generic model control (GMC). *Comp. Chem. Eng.*, 12:573–580, 1988.

H. Nijmeijer and A.J. VanderSchaft. *Nonlinear Dynamical Systems*. Springer-Verlag, New York, 1990.

S. Sastry and M. Bodson. *Adaptive Control - Stability, Convergence, and Robustness*. Prentice Hall, New Jersey, 1989.

J.J.E. Slotine and W. Li. *Applied Nonlinear Control*. Prentice Hall, Englewood Cliffs, 1991.

M. Soroush and C. Kravaris. Nonlinear control of a batch polymerization reactor: an experimental study. *AIChE J.*, 38(9):1429–1448, 1992.

M. Soroush and C. Kravaris. Nonlinear control of a polymerization CSTR with singular characteristic matrix. *AIChE J.*, 40(6):980–990, 1994.

R. Su. On the linear equivalents of nonlinear systems. *Syst. Control Lett.*, 2: 48–52, 1982.

L.C. To, M.O. Tadé, M. Kraetzl, and G.P. Le Page. Nonlinear control of a simulated industrial evaporator process. *J. Proc. Cont.*, 5(3):173–182, 1995.

M. Vidyasagar. *Nonlinear Systems Analysis*. Prentice Hall, Englewood Cliffs, 1993.

Z.L. Wang, F. Pla, and J.P. Corriou. Nonlinear adaptive control of batch styrene polymerization. *Chem. Eng. Sci.*, 50(13):2081–2091, 1995.

W.M. Wohnam. *Linear Multivariable Control. A Geometric Approach*. Springer-Verlag, New York, 1985.

Chapter 18

State Observers

18.1 Introduction

In many processes, e.g. in fine chemistry, biotechnology, measurements are difficult, needless to say scarce. Only some variables are easily measurable: temperature, pressure, flow rate, pH and concentration of some species. Nevertheless, to optimally control the processes in spite of the absence of adequate physical sensors, it is frequently necessary to know the values of variables that are not directly available to measurement.

Thus, in the biological processes of fermentation, the biomass is difficult to measure on-line either for reasons of sampling difficulties or potential contamination problems, but it can be estimated from on-line measurements of pH, oxygen and carbon dioxide, and also from the studied fermentation process model. The stoichiometric relations, mass balances and possibly heat balances will be used. Given the low quality of the measurements, moreover it is generally necessary to filter them before use.

In a distillation column, the objective is to regulate top and bottom mole fractions; in the case of nonlinear control of this column, the vapour and liquid mole fractions inside the column need to be known; they can be reconstructed from the temperature measurements on sensitive trays and from the knowledge model of column. Besides the control itself, the prediction of the mole fractions profile in the column can allow us to realize not only monitoring for detection of misfunctioning, but also failure diagnostic (Li and Olson, 1991).

An observer, or state estimator, or again intelligent sensor or soft sensor, is an algorithm based on the knowledge of models describing the behaviour of the process and using measurements acquired in the process to reconstruct the missing measurements. The possibilities of observers can be extended to the estimation of process characteristic parameters, such as in chemical engineering: heat of reaction, global heat-transfer or mass-transfer coefficient and kinetic constants.

The observer is most often designed to estimate the states in a closed-loop control system needing the knowledge of the states; it can also simply be used

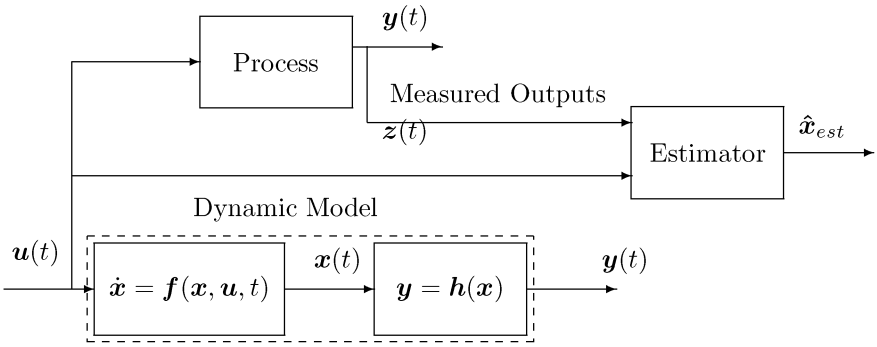


Figure 18.1: Observer principle

for process monitoring or diagnostics.

18.1.1 Indirect Sensors

Many physical sensors are based on a static model such as a mercury thermometer that uses the relation between quasi-linear dilatation of the mercury (observed phenomenon) and temperature, or a pH-meter that uses the potential difference between two electrodes and the electrochemistry laws to provide an indication of the H^+ concentration. These are indirect sensors.

18.1.2 Observer Principle

An observer (Luenberger, 1966, 1971; Misawa and Hedrick, 1989), can be described by the diagram in Fig. 18.1. The physical sensor realizes measurements in the process. The known control inputs and part or all of the measurements constitute the inputs of the estimator which gives an estimation of the states (eventually parameters) which are useful for monitoring or control. The estimator is developed from a linear or nonlinear dynamic model of the process. This state estimator can be deterministic such as the Luenberger observer or stochastic such as the Kalman filter, depending on whether the process model is deterministic or stochastic.

The most general dynamic model is written in state space as

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}, u, \theta, t) \\ \boldsymbol{y} &= \boldsymbol{h}(\boldsymbol{x}, t) \end{aligned} \tag{18.1}$$

where $\boldsymbol{x}$ represents the state vector and θ the parameter vector on which the model depends.

The general equation of an observer is resumed as

$$\begin{aligned} \text{new estimated state} = \text{old estimated state} + \\ \text{gain} \times [\ \text{measurement} - \text{measurement estimation}\ ] \end{aligned}$$

which means that a state estimation is realized from the previous state estimation by adding a correction that is proportional to the deviation between

the measurement and its estimation. The objective of an observer consists of calculating in an optimal manner the gain that is the proportionality factor, while ensuring the estimation stability.
Another expression is possible:

$$\text{derivative of the estimated state} = \text{dynamic state model} + \text{gain} \times [\ \text{measurement} - \text{measurement estimation}\ ]$$

This second form can be considered as a continuous form (making use of the process continuous model) whereas the first form is discrete (uses time $t - \Delta t$ and time t).

The observer convergence and stability are key-points. Two types of convergence are theoretically defined: exponential or asymptotic convergence. The quality of estimation will depend amongst other things on the model and measurement quality.

18.2 Parameter Estimation

In many cases of processes, the parameters are ill-known (heat of reaction, transfer coefficient, etc.), progressively time-varying (catalyst activity, fouling, etc.) or fast (leak in a pressure vessel, choking in a column, etc.). The engineer responsible for the process might wish to evaluate them or display a change of behaviour. The particular case of biological reactors with the estimation of kinetic parameters has been studied (Bastin and Dochain, 1990; Cheruy and Flaus, 1994). In this perspective, it is possible to consider an augmented state vector constituted by the model states $\boldsymbol{x}$ and the parameters θ concerned by the estimation procedure. The linear or extended Kalman filter (Ljung, 1979) will be applied to this augmented state vector in an analogous way to the estimation of the states alone. There exists a recursive form of the linear Kalman filter used for parameter estimation which needs no matrix inversion; it is based on the matrix inversion lemma (Söderström and Stoica, 1989).

18.3 Statistical Estimation

In existing processes, due to process computers connected to numerous sensors with small sampling periods, large amounts of data are collected. For process monitoring, fault detection and statistical process control, statistical methods such as PCA (principal component analysis) or PLS (partial least squares or projection to latent structures) can be implemented to reduce the information so that simple models retain a maximum of information (Burnham et al., 1996; Kaspar and Ray, 1992; Kresta et al., 1991; Rencher, 1995). This type of method has been used, in particular, for distillation columns (Kano et al., 2000; Kresta et al., 1991; Mejdell and Skogestad, 1991a,b) to estimate the product compositions from the temperature measurements along the column.

The objective of these methods is thus to provide a predictive model that is capable of explaining the main influences of the process measured variables

on the product quality variables. The dimension of the model is very much reduced compared to the dimension of the concerned data group.

18.3.1 About the Data

PCA (principal component analysis) and PLS (partial least squares) are designed to treat large amounts of correlated data (Otto, 1999; Tenenhaus, 1998). Following Kresta et al. (1991), the data are classified in two groups:

- The process measurements or inputs grouped in a matrix $\boldsymbol{X}$ (size $n \times k$) containing k variables, each with n observations. As k is larger than 1, the analysis is called multivariate.
- The product quality variables or outputs (of number m) grouped in a matrix $\boldsymbol{Y}$ (size $n \times m$).

The data $\boldsymbol{X}$ are, in general, transformed into reduced deviation variables. The data $\boldsymbol{Y}$ are often only centred. Consider, for example, the matrix $\boldsymbol{X}$. They are first centred around the mean of the calibration set. Then, the centred data are scaled to unit variance.

Finally, the matrix $\boldsymbol{X}$ is transformed into a matrix $\boldsymbol{Z}$ such that

$$\boldsymbol{Z} = \left[\frac{x_{ij} - \bar{x}_j}{s_j}\right] \quad \text{with:} \quad \bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \quad \text{and:} \quad s_j = \sqrt{\frac{\sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}{n-1}} \tag{18.2}$$

The covariance matrix for the data matrix $\boldsymbol{X}$ is equal to

$$\text{Cov}(\boldsymbol{X}) = [\text{Cov}(i,j)]$$
$$\text{with} \begin{cases} \text{Cov}(i,i) = \dfrac{1}{n-1}\displaystyle\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \qquad i = 1, \ldots, p \\ \text{Cov}(i,j) = \dfrac{1}{n-1}\displaystyle\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)\,(x_{kj} - \bar{x}_j) \qquad i,j = 1, \ldots, p \quad ; \quad i \neq j \end{cases} \tag{18.3}$$

If the data matrix $\boldsymbol{X}$ is centred and scaled to unit variance, the covariance matrix of $\boldsymbol{X}$ is equal to the correlation matrix of $\boldsymbol{X}$ defined as

$$\text{Corr}(\boldsymbol{X}) = [\text{Corr}(i,j)]$$
$$\text{with} \begin{cases} \text{Corr}(i,i) = 1 \qquad i = 1, \ldots, p \\ \text{Corr}(i,j) = \dfrac{\text{Cov}(i,j)}{s_i\, s_j} \qquad i,j = 1, \ldots, p \quad ; \quad i \neq j \end{cases} \tag{18.4}$$

where s_i (similarly for s_j) is the standard deviation defined in (18.2).

18.3.2 Principal Component Analysis

In principal component analysis (PCA), data are modelled using orthogonal components in order of decreasing importance. The dependent data are thus transformed into significant independent ones.

According to singular value decomposition (SVD), the matrix $\boldsymbol{X}$ can be decomposed into

$$\boldsymbol{X} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T \tag{18.5}$$

where $\boldsymbol{U}$ is the matrix formed by the normalized eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$ ordered as columns, $\boldsymbol{V}$ is the matrix formed by the normalized eigenvectors of the correlation matrix $\boldsymbol{X}^T\boldsymbol{X}$ ordered as columns and $\boldsymbol{\Sigma}$ is the diagonal matrix having its elements equal to the square roots (singular values of $\boldsymbol{X}$) of the ordered eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$.

In PCA, the matrix $\boldsymbol{X}$ is decomposed into

$$\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{t}_i \boldsymbol{p}_i^T + \boldsymbol{E}_N \tag{18.6}$$

where N is the chosen or optimal number of latent vectors, $\boldsymbol{t}_i$ are the scores or principal components associated with the i-th latent vector of $\boldsymbol{X}$, $\boldsymbol{p}_i$ is the loading vector for the i-th latent vector of $\boldsymbol{X}$ and $\boldsymbol{E}_N$ is the residual matrix of $\boldsymbol{X}$ due to the reduction of dimensionality. $\boldsymbol{X}$ is projected onto a reduced N-dimension subspace by the projection matrix $\boldsymbol{P}$ and takes the coordinates $\boldsymbol{T}$ in this subspace. The $\boldsymbol{t}_i$'s and $\boldsymbol{p}_i$'s are forced to be orthogonal, and orthogonal to the rows or columns of $\boldsymbol{E}$ so that $\boldsymbol{t}_i$'s are eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{p}_i$'s are eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$, as in SVD. It results that SVD and PCA are related by $\boldsymbol{P} = \boldsymbol{V}$ and $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Sigma}$ if $\boldsymbol{E} = 0$. Thus, the data contained in $\boldsymbol{X}$ are modelled using orthogonal components. As for SVD decomposition where the eigenvalues are calculated in decreasing magnitude order, in PCA the eigenvectors are calculated in order of decreasing importance. However, SVD is inefficient at calculating the principal components for correlated data sets with $N \ll k$. SVD would be equivalent to PCA if the contribution of the singular values in $\boldsymbol{\Sigma}$ lower than a given threshold were neglected. The optimal dimension N can be obtained by verifying that the addition of new principal components adds little information, or it can be done by cross-validation on another set of data for which the prediction error sum of squares is calculated. When the latter value is sufficiently small, the number N is optimal. Often, a percentage of the explained variance is given as

$$100\,\frac{\sum_{i,j} X_{ij}^2 - \sum_{i,j}(X_{ij} - t_i p_j)^2}{\sum_{i,j} X_{ij}^2} \tag{18.7}$$

After retaining a few principal components $(2, 3, \ldots)$, the data are represented graphically in the planes formed by two given principal components.

The matrix $\boldsymbol{Y}$ is similarly decomposed into

$$\boldsymbol{Y} = \sum_{i=1}^{N} \boldsymbol{u}_i \boldsymbol{q}_i^T + \boldsymbol{F}_N \tag{18.8}$$

The algorithm describing the iterative principal component analysis called NIPALS (nonlinear iterative partial least squares) working on the data $\boldsymbol{X}$ is the following:

Initially, centre and scale to unit variance the data matrix $\boldsymbol{X}$. Note $\boldsymbol{X}_0 = \boldsymbol{X}$.

N being the number of principal components, denote by k an iteration step with $k = 1, \ldots, N$.

1. Set $\boldsymbol{t_k}$ equal to a column of $\boldsymbol{X}_{k-1}$, in general the first one.
2. Calculate the loading vector $\boldsymbol{p}_k = \boldsymbol{X}_{k-1}^T \boldsymbol{t}_k / \boldsymbol{t}_k^T \boldsymbol{t}_k$
3. Normalize the loading vector $\boldsymbol{p}_k = \boldsymbol{p}_k / \|\boldsymbol{p}_k\|$.
4. Calculate the scores $\boldsymbol{t}_k = \boldsymbol{X}_{k-1} \boldsymbol{p}_k / \boldsymbol{p}_k^T \boldsymbol{p}_k$.
5. Compare new and old $\boldsymbol{t}$ vectors: if convergence, go to step 6, else go to step 2.
6. Calculate the residual matrix $\boldsymbol{E}_k = \boldsymbol{X}_k = \boldsymbol{X}_{k-1} - \boldsymbol{t}_k \boldsymbol{p}_k^T$. If the number of principal components is sufficient, end, else go to step 1.

Finally, the principal component prediction model is

$$\hat{\boldsymbol{X}} = \boldsymbol{T}\, \boldsymbol{P}^T \tag{18.9}$$

where $\boldsymbol{T}$ and $\boldsymbol{P}$ represent the storage matrices for the successive vectors $\boldsymbol{t}$ and $\boldsymbol{p}$.

After convergence of the sequence of steps 2 to 5, the vectors $\boldsymbol{p}$ and $\boldsymbol{t}$ tend respectively towards the eigenvectors of the respective matrices $\frac{1}{n-1} \boldsymbol{X}^T \boldsymbol{X}$ and $\frac{1}{n-1} \boldsymbol{X} \boldsymbol{X}^T$. Furthermore, the corresponding eigenvalue is equal to $\frac{1}{n-1} \boldsymbol{t}^T \boldsymbol{t}$. Thus, an often used index is the fraction of explained variance, defined as

$$f = \frac{\sum_{i=1}^{N} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{18.10}$$

where the λ_i's are the eigenvalues and N is the number of principal components. The numerator considers only the larger eigenvalues retained in the PCA and the denominator considers all the eigenvalues of the matrix $\frac{1}{n-1} \boldsymbol{X}^T \boldsymbol{X}$.

Note that the NIPALS algorithm described above can be used with little difference in the case of missing data (Tenenhaus, 1998).

18.3.3 Partial Least Squares

Partial least squares (PLS) is a method of analysis of data developed in particular by Wold et al (Trygg and Wold, 2002; Wold et al., 1984, 1987, 2001), and many reviews are available (Geladi, 1988; Geladi and Kowalski, 1986; Höskuldsson, 1988). The PLS method presents many common points with the PCA method but it works simultaneously on the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. The loading vectors which were orthogonal in PCA are not orthogonal anymore in PLS.

The PLS algorithm (Kresta et al., 1991; Tenenhaus, 1998) is the following:

Initially, centre and scale to unit variance the data of $\boldsymbol{X}$, centre (and in general scale) the data of $\boldsymbol{Y}$. Note $\boldsymbol{X}_0 = \boldsymbol{X}$ and $\boldsymbol{Y}_0 = \boldsymbol{Y}$.

N being the number of principal components, denote by k an iteration step with $k = 1, \ldots, N$.

1. Set $\boldsymbol{u}_k$ equal to a column of $\boldsymbol{Y}$, in general the first.

2. Calculate $\boldsymbol{w}_k = \boldsymbol{X}_{k-1}^T \boldsymbol{u}_k / \boldsymbol{u}_k^T \boldsymbol{u}_k$ (the columns of $\boldsymbol{X}$ are regressed on $\boldsymbol{u}$).
3. Normalize $\boldsymbol{w}_k$ so that $|\boldsymbol{w}_k| = 1$.
4. Calculate the scores $\boldsymbol{t}_k = \boldsymbol{X}_{k-1} \boldsymbol{w}_k / \boldsymbol{w}_k^T \boldsymbol{w}_k$.
5. Calculate $\boldsymbol{q}_k^T = \boldsymbol{t}_k^T \boldsymbol{Y} / \boldsymbol{t}_k^T \boldsymbol{t}_k$ (the columns of $\boldsymbol{Y}$ are regressed on $\boldsymbol{t}$).
6. Calculate new vector $\boldsymbol{u}_k = \boldsymbol{Y} \boldsymbol{q}_k / \boldsymbol{q}_k^T \boldsymbol{q}_k$.
7. Compare new and old $\boldsymbol{u}$ vectors: if convergence, go to step 8, else go to step 2.
8. Calculate the $\boldsymbol{X}$ loadings $\boldsymbol{p}_k = \boldsymbol{X}^T \boldsymbol{t}_k / \boldsymbol{t}_k^T \boldsymbol{t}_k$.
9. Calculate the residuals $\boldsymbol{X}_k = \boldsymbol{X}_{k-1} - \boldsymbol{t}_k \boldsymbol{p}_k^T$ and $\boldsymbol{Y}_k = \boldsymbol{Y}_{k-1} - \boldsymbol{t}_k \boldsymbol{q}_k^T$. Increment k by 1. If $k \leq N$, go to step 1, else end.

This algorithm is a modified version of the original PLS algorithm, where in steps 1, 5, 6, 8, the matrices $\boldsymbol{X}_{k-1}$ and $\boldsymbol{Y}_{k-1}$ are used instead of respectively $\boldsymbol{X}$ and $\boldsymbol{Y}$ as here.

The predicted response matrix by the PLS model is $\hat{\boldsymbol{Y}} = \boldsymbol{T}\,\boldsymbol{Q}^T$, where $\boldsymbol{T}$ and $\boldsymbol{Q}$ represent the storage matrices for the successive vectors $\boldsymbol{t}_k$ and $\boldsymbol{q}_k$, e.g.

$$\boldsymbol{T} = [\boldsymbol{t}_1 \quad \boldsymbol{t}_2 \quad \ldots \quad \boldsymbol{t}_N] \tag{18.11}$$

For the whole matrix of data $\boldsymbol{X}$, the corresponding predicted matrix can be given as

$$\hat{\boldsymbol{Y}} = \sum_{k=1}^{N} \boldsymbol{t}_k \, \boldsymbol{q}_k^T = \boldsymbol{T}\,\boldsymbol{Q}^T \qquad \text{with:} \quad \boldsymbol{T} = \boldsymbol{X}\,\boldsymbol{W}(\boldsymbol{P}^T\,\boldsymbol{W})^{-1} \tag{18.12}$$

where $\boldsymbol{P}$ and $\boldsymbol{W}$ are respectively the storage matrices for the successive vectors $\boldsymbol{p}_k$ and $\boldsymbol{w}_k$.

The regression model with respect to centred and scaled data gives the row vector of predicted variables

$$\hat{\boldsymbol{y}} = \boldsymbol{x}\,\boldsymbol{W}\,(\boldsymbol{P}^T\,\boldsymbol{W})^{-1}\,\boldsymbol{Q}^T \tag{18.13}$$

where $\boldsymbol{x}$ is the analytical row vector of components $x_1, \ldots, x_{n_x}$.

Algorithms exist for nonlinear PLS (Hassel et al., 2002) and dynamic PLS (Lakshminarayanan et al., 1997).

18.4 Observers

18.4.1 Luenberger Observer

Consider a system described by the linear deterministic model

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\,\boldsymbol{x}(t) + \boldsymbol{B}\,\boldsymbol{u}(t) \\ \boldsymbol{y}(t) &= \boldsymbol{C}\,\boldsymbol{x}(t) \end{aligned} \tag{18.14}$$

A Luenberger observer (Luenberger, 1966, 1971, 1979), is described by the following dynamic system

$$\begin{aligned} \dot{\boldsymbol{z}} &= \boldsymbol{A}\,\boldsymbol{x}(t) + \boldsymbol{B}\,\boldsymbol{u}(t) + \boldsymbol{G}\,(\boldsymbol{y}(t) - \boldsymbol{C}\,\boldsymbol{x}(t)) \\ &= (\boldsymbol{A} - \boldsymbol{G}\,\boldsymbol{C})\,\boldsymbol{x}(t) + \boldsymbol{B}\,\boldsymbol{u}(t) + \boldsymbol{G}\,\boldsymbol{y}(t) \end{aligned} \tag{18.15}$$

The dynamic response is determined by the matrix $\boldsymbol{A} - \boldsymbol{GC}$ whose eigenvalues can be fixed (provided the system is observable) by conveniently choosing the matrix $\boldsymbol{G}$.

18.4.2 Linear Kalman Filter

The linear Kalman filter (Kalman, 1960; Kalman and Bucy, 1961) has been described in Chap. 11. This optimal filter is a method of stochastic estimation that calculates the gain matrix by minimization of the sum of the estimation error variances (Bozzo, 1983).

In general, two types of linear Kalman filters are distinguished: the continuous-discrete Kalman filter where the model is continuous and the measurements are discrete, the discrete-discrete Kalman filter where the model is discrete and the measurements are discrete.

An implementation of the discrete linear Kalman filter (Brown and Hwang, 1997) easily usable for state estimation is proposed in Eqs. (11.114) to (11.116).

In Gaussian linear quadratic control (Sect. 14.6.2) that uses the linear Kalman filter to estimate the states, the separation principle based on the closed-loop model eigenvalues was shown: it is possible to separately determine the observer and the state feedback optimal control law. Similarly, the separation principle is applicable to Luenberger observer.

18.4.3 Extended Kalman Filter (EKF) in Continuous-Discrete Form

The extended Kalman filter (EKF) is an extension of the linear Kalman filter to the case where the system is described in state space in a nonlinear form. The derivative of the state vector is equal to

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, u, \theta, t) + \boldsymbol{w} \tag{18.16}$$

where $\boldsymbol{w}$ is a Gaussian noise of zero mean and covariance matrix $\boldsymbol{Q}$. The measured outputs z are described by the model

$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{v}_k \tag{18.17}$$

where $\boldsymbol{v}_k$ is a Gaussian noise of zero mean and covariance matrix $\boldsymbol{R}_k$.

In this way, EKF is defined in the continuous-discrete form, i.e. the process is continuous and the outputs are discrete as they are measured at sampling times denoted by k, which are in general regularly, but not necessarily, spaced.

In this first approach, the model parameters θ are assumed to be perfectly known; the model can thus be written in simplified form

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, u, t) + \boldsymbol{w} \tag{18.18}$$

The problem is to estimate at time k the unmeasured states by using the available process measurements at time $k-1$. The recurrent algorithm of the filter includes two stages:

1. Propagation of the state estimation and the error covariance matrix.
This means that the differential equations describing the variation of the state vector $\boldsymbol{x}$ and the error covariance matrix $\boldsymbol{P}$ are integrated in the time interval $[k-1, k]$ to obtain a prediction, whereas the measurement has not yet been realized. The predictions of $\boldsymbol{x}(t_k)$ and $\boldsymbol{P}(t_k)$ are respectively denoted by $\hat{\boldsymbol{x}}_k(-)$ and $\boldsymbol{P}_k(-)$. The differential equations to be integrated are

$$\dot{\hat{\boldsymbol{x}}}(t) = \boldsymbol{f}(\hat{\boldsymbol{x}}, u, t) \qquad ; \quad \hat{\boldsymbol{x}}(0) = \boldsymbol{x}_0 \tag{18.19}$$

$$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(\hat{\boldsymbol{x}}, t)\, \boldsymbol{P}(t) + \boldsymbol{P}(t)\, \boldsymbol{F}^t(\hat{\boldsymbol{x}}, t) + \boldsymbol{Q}(t) \qquad ; \quad \hat{\boldsymbol{P}}(0) = \boldsymbol{P}_0 \tag{18.20}$$

where $\boldsymbol{F}$ is the Jacobian matrix of $\boldsymbol{f}$ equal to

$$\boldsymbol{F}(\hat{\boldsymbol{x}}, t) = \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \right)_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \tag{18.21}$$

2. Update of the state estimation and the error covariance matrix.
The measurements realized at instant t_k are used to correct the estimations $\hat{\boldsymbol{x}}_k(-)$ and $\boldsymbol{P}_k(-)$ by minimizing the estimation error. The estimations thus corrected, obtained at instant t_k, are denoted by $\hat{\boldsymbol{x}}_k(+)$ and $\boldsymbol{P}_k(+)$ and are equal to

$$\hat{\boldsymbol{x}}_k(+) = \hat{\boldsymbol{x}}_k(-) + \boldsymbol{K}_k \left[\boldsymbol{z}_k - \boldsymbol{h}(\hat{\boldsymbol{x}}_k(-)) \right] \tag{18.22}$$

$$\boldsymbol{P}_k(+) = \left[\boldsymbol{I} - \boldsymbol{K}_k\, \boldsymbol{H}_k(\hat{\boldsymbol{x}}_k(-)) \right]\, \boldsymbol{P}_k(-) \tag{18.23}$$

where $\boldsymbol{H}_k$ is the Jacobian matrix of $\boldsymbol{h}$ equal to

$$\boldsymbol{H}_k(\hat{\boldsymbol{x}}_k(-)) = \left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} \right)_{\boldsymbol{x}=\hat{\boldsymbol{x}}_k(-)} \tag{18.24}$$

and $\boldsymbol{K}_k$ is the Kalman gain matrix equal to

$$\boldsymbol{K}_k = \boldsymbol{P}_k(-)\, \boldsymbol{H}_k^T(\hat{\boldsymbol{x}}_k(-)) \left[\boldsymbol{H}_k(\hat{\boldsymbol{x}}_k(-))\, \boldsymbol{P}_k(-)\, \boldsymbol{H}_k^T(\hat{\boldsymbol{x}}_k(-)) + \boldsymbol{R}_k \right]^{-1} \tag{18.25}$$

Very frequently, the noise covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}_k$ that represent the measurement of the model and measurement uncertainty are assumed to be diagonal. Their adjustment can be done by simulation.

The Kalman gain $\boldsymbol{K}$ can affect in different ways the various estimated states. Indeed, the elements of the Jacobian matrix $\boldsymbol{F}$ represent, through the model, the sensitivity of the different states one to each other.

Because of its similarity with the linear Kalman filter, the extended Kalman filter (Fig. 18.2) is often used although it presents some drawbacks:

- P being only an approximation of the true covariance matrix, the extended Kalman filter performance cannot be guaranteed and its stability cannot be proved.
- The extended Kalman filter equations assume that the process model is exact. No robustness is guaranteed against modelling errors.

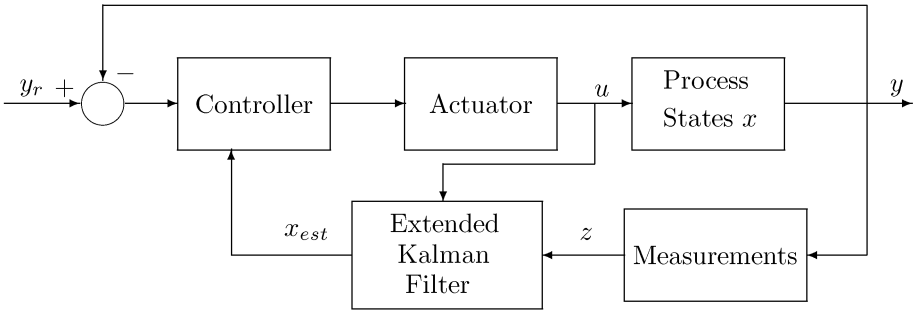


Figure 18.2: Implementation of the extended Kalman filter in a control scheme

For these reasons, variants have been designed (Misawa and Hedrick, 1989) such as the constant gain extended Kalman filter (Safonov and Athans, 1978) which was essentially proposed to avoid the long calculations related to the update of the state and covariance matrix estimations. Becerra et al. (2001) have applied the extended Kalman filter to systems described by differential-algebraic models.

Industrial experiences related to the use of extended Kalman filter are cited, in particular, in the article of (Wilson et al., 1998) where general recommendations are given and cautions are presented for industrial applications concerning difficulties that scarcely appear in simulation.

18.4.4 Globally Linearizing Observer

The development of the globally linearizing observer (Isidori, 1995; Krener and Isidori, 1983) can be considered dually by comparison with the exact state-space linearization which allowed us to build a state feedback with prescribed eigenvalues. In the case of the observer synthesis, we are looking for an observation error dynamics which becomes, after coordinate change, linear and whose eigenvalues can be prescribed.

Consider the nonlinear system

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}) \\ y &= h(\boldsymbol{x}) \end{aligned} \tag{18.26}$$

We seek a diffeomorphism $z = \Phi(x)$ transforming the system (18.26) into a linear system

$$\begin{aligned} \dot{\boldsymbol{z}} &= \left[\frac{\partial \Phi}{\partial \boldsymbol{x}}\, \boldsymbol{f}(\boldsymbol{x})\right]_{\boldsymbol{x}=\Phi^{-1}(\boldsymbol{z})} &&= \boldsymbol{A}\,\boldsymbol{z} + \boldsymbol{k}(\boldsymbol{y}) = \boldsymbol{A}\,\boldsymbol{z} + \boldsymbol{k}(\boldsymbol{C}\,\boldsymbol{z}) \\ y &= h(\Phi^{-1}(\boldsymbol{z})) &&= \boldsymbol{C}\,\boldsymbol{z} \end{aligned} \tag{18.27}$$

where the pair $(\boldsymbol{A}, \boldsymbol{C})$ is observable and $\boldsymbol{k}$ is a function vector.

In these conditions, the observer is given by

$$\begin{aligned}\dot{\boldsymbol{\xi}} &= \boldsymbol{A}\,\boldsymbol{\xi} + \boldsymbol{k}(\boldsymbol{C}\,\boldsymbol{\xi}) + \boldsymbol{G}\,(\boldsymbol{C}\,\boldsymbol{\xi} - y)\\ &= (\boldsymbol{A} + \boldsymbol{G}\,\boldsymbol{C})\,\boldsymbol{\xi} + \boldsymbol{k}(y) - \boldsymbol{G}\,y\end{aligned} \tag{18.28}$$

The observation error, i.e. the difference between the theoretical and estimated state, is equal to

$$\boldsymbol{e} = \boldsymbol{\xi} - \boldsymbol{z} = \boldsymbol{\xi} - \Phi(\boldsymbol{x}) \tag{18.29}$$

and its dynamics is described by the linear system

$$\dot{\boldsymbol{e}} = (\boldsymbol{A} + \boldsymbol{G}\,\boldsymbol{C})\,\boldsymbol{e} \tag{18.30}$$

whose eigenvalues can be prescribed through the gain vector $\boldsymbol{G}$.

The observer linearization is possible if and only if

$$\dim(\text{span}\left\{dh(\boldsymbol{x}^\circ), dL_{\boldsymbol{f}}h(\boldsymbol{x}^\circ), \ldots, dL_{\boldsymbol{f}}^{n-1}h(\boldsymbol{x}^\circ)\right\}) = n \tag{18.31}$$

We then seek the unique vector $\boldsymbol{\tau}$ in the neighbourhood U of $\boldsymbol{x}^\circ$ such that

$$\begin{aligned}&L_{\boldsymbol{\tau}}h(\boldsymbol{x}) = L_{\boldsymbol{\tau}}L_{\boldsymbol{f}}h(\boldsymbol{x}) = \ldots = L_{\boldsymbol{\tau}}L_{\boldsymbol{f}}^{n-2}h(\boldsymbol{x}) = 0\\ &L_{\boldsymbol{\tau}}L_{\boldsymbol{f}}^{n-1}h(\boldsymbol{x}) = 1\end{aligned} \tag{18.32}$$

then the mapping $\boldsymbol{F}$ such that its Jacobian matrix is

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{z}} = [\boldsymbol{\tau}(\boldsymbol{x}) \quad -\text{ad}_{\boldsymbol{f}}\boldsymbol{\tau}(\boldsymbol{x}) \quad \ldots \quad (-1)^{n-1}\text{ad}_{\boldsymbol{f}}^{n-1}\boldsymbol{\tau}(\boldsymbol{x})]_{\boldsymbol{x}=\boldsymbol{F}(\boldsymbol{z})} \tag{18.33}$$

providing a system of n partial differential equations. Then we set

$$\Phi = \boldsymbol{F}^{-1} \tag{18.34}$$

and we use

$$\boldsymbol{k}(\boldsymbol{z}_n) = \left[\frac{\partial \Phi}{\partial \boldsymbol{x}}\,\boldsymbol{f}(\boldsymbol{x})\right]_{\boldsymbol{x}=\Phi^{-1}(\boldsymbol{z})} - \begin{bmatrix} 0\\ z_1\\ \vdots\\ z_{n-1}\end{bmatrix} \tag{18.35}$$

The implementation of such an observer is, of course, not very simple and the following high-gain observer is easier to implement.

18.4.5 High-Gain Observer

The high-gain observer (Gauthier et al., 1992) is a kind of extended Luenberger observer. Consider the nonlinear system

$$\begin{aligned}\dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,u\\ y &= h(\boldsymbol{x})\end{aligned} \tag{18.36}$$

The mapping Φ defined by

$$\Phi(\boldsymbol{x}) = \begin{bmatrix} h(\boldsymbol{x}) \\ L_f h(\boldsymbol{x}) \\ \vdots \\ L_f^{n-1} h(\boldsymbol{x}) \end{bmatrix} \tag{18.37}$$

constitutes a diffeomorphism that transforms the system (18.36) into

$$\begin{aligned} \dot{\boldsymbol{x}} &= \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \\ \phi(\boldsymbol{x}) \end{bmatrix} + \begin{bmatrix} g_1(x_1) \\ g_2(x_1, x_2) \\ \vdots \\ \vdots \\ g_n(x_1, \ldots, x_n) \end{bmatrix} u = \boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{G}(\boldsymbol{x})\, u \\ y &= x_1 = \boldsymbol{C}\, \boldsymbol{x} \end{aligned} \tag{18.38}$$

The system is assumed to be uniformly observable.

The following dynamic system

$$\dot{\hat{\boldsymbol{x}}} = \boldsymbol{F}(\hat{\boldsymbol{x}}) + \boldsymbol{G}(\hat{\boldsymbol{x}})\, u - \boldsymbol{S}_\infty^{-1} \boldsymbol{C}^T (\boldsymbol{C}\, \hat{\boldsymbol{x}} - y) \tag{18.39}$$

constitutes an observer for the system (18.36). The positive definite matrix $\boldsymbol{S}_\infty$ is the solution of the equation

$$-\theta\, \boldsymbol{S}_\infty - \boldsymbol{A}^T\, \boldsymbol{S}_\infty - \boldsymbol{S}_\infty\, \boldsymbol{A} + \boldsymbol{C}^T\, \boldsymbol{C} = 0 \tag{18.40}$$

where θ is a large scalar that allows us to adjust the convergence rate of the estimator, with $\boldsymbol{C} = [1\, 0 \ldots 0]$ and $\boldsymbol{A}$ is the matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ \vdots & & \ldots & 0 & 1 \\ 0 & \ldots & & \ldots & 0 \end{bmatrix} \tag{18.41}$$

The observer thus defined converges exponentially. It has been applied in simulation in biological reactors (Gauthier et al., 1992).

A slightly different version of the observer, where the matrix $\boldsymbol{S}_\infty$ is replaced by matrix $\boldsymbol{S}$ calculated as the solution of the differential equation

$$\dot{\boldsymbol{S}} = -\theta\, \boldsymbol{S} - \boldsymbol{A}^T\, \boldsymbol{S} - \boldsymbol{S}\, \boldsymbol{A} + \boldsymbol{C}^T\, \boldsymbol{C} \tag{18.42}$$

has also been used in simulation in chemical reactors (Gibon-Fargeot et al., 1994).

18.4.6 Moving state estimation

The moving state observer is an efficient means of estimating the states, with in particular the possibility to constrain the states, ouputs and noises. It can be grossly described as a least-squares optimization leading to an estimation of the states and working with a limited amount of information. It avoids the recursive manner characteristic of the extended Kalman filter. It possesses common characters with model predictive control (chapter 16). It has been studied by several researchers (Alamir, 1999; Kwon et al., 1999; Michalska and Mayne, 1995; Rao et al., 2001; Robertson et al., 1996; Slotine et al., 1987; Wang et al., 1997; Zimmer, 1994) under different approaches, however presenting some similarities.

The full and moving state estimations follow more or less the same steps. The main difference of the moving state estimation with respect to the full state estimation resides in the handling of the variables. In the full state estimation, at current time k, all variables (manipulated inputs, measured outputs, estimated states) from initial time $i = 0$ to $i = k$ are necessary and used in the calculation. In the moving state estimation with horizon H, only the concerned variables (manipulated inputs, measured outputs, estimated states) from $i = k + 1 - H$ to $i = k$ are necessary and used in the calculation, they are collected in moving vectors.

First consider the problem of full state estimation. Assume that the process is represented by the following continuous-time model similar to the model used by the extended Kalman filter

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t)) + \boldsymbol{G}\boldsymbol{w}(t) \tag{18.43}$$

where $\boldsymbol{w}$ is a Gaussian noise of zero mean. The measured outputs y are described by the discrete-time model

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{v}_k \tag{18.44}$$

where $\boldsymbol{v}_k$ is a Gaussian noise of zero mean.

The continuous nonlinear model (18.43) is approximated by the linear discrete model

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\,\boldsymbol{x}_k + \boldsymbol{B}\,\boldsymbol{u}_k + \boldsymbol{G}\,\boldsymbol{w}_k \tag{18.45}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the Jacobian matrices with respect to $\boldsymbol{x}_k$ and $\boldsymbol{u}_k$ respectively. The measurement model is linearized as

$$\boldsymbol{y}_{k+1} = \boldsymbol{C}\,\boldsymbol{x}_{k+1} + \boldsymbol{v}_{k+1} \tag{18.46}$$

where $\boldsymbol{C}$ is the Jacobian matrix of $\boldsymbol{h}$ with respect to $\boldsymbol{x}_k$.

In the full state estimation problem, the following criterion

$$J_k = (\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0)^T\,\boldsymbol{\Pi}_0^{-1}\,(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0) + \sum_{i=0}^{k-1}(\boldsymbol{v}_{i+1}^T\,\boldsymbol{R}^{-1}\,\boldsymbol{v}_{i+1} + \boldsymbol{w}_i^T\,\boldsymbol{Q}^{-1}\,\boldsymbol{w}_i) \tag{18.47}$$

is minimized with respect to the initial state $\boldsymbol{x}_0$ and to the sequence of noises $\{\boldsymbol{w}_0, \ldots, \boldsymbol{w}_{k-1}\}$. Then the states $\hat{\boldsymbol{x}}_i$ are obtained by use of equation (18.45).

The weighting matrices $\boldsymbol{Q}^{-1}$, $\boldsymbol{R}^{-1}$ and $\boldsymbol{\Pi}^{-1}$ respectively symbolize the confidence in the dynamic model, the measurements and the initial estimation.

For example, at time $k = 1$ when the first optimization problem is solved in full state estimation, the criterion J_1

$$J_1 = (\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0)^T \boldsymbol{\Pi}_0^{-1} (\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0) + (\boldsymbol{v}_1^T \boldsymbol{R}^{-1} \boldsymbol{v}_1 + \boldsymbol{w}_0^T \boldsymbol{Q}^{-1} \boldsymbol{w}_0) \tag{18.48}$$

is minimized with respect to $\boldsymbol{x}_0$ and $\boldsymbol{w}_0$ based on the available measurement $\boldsymbol{y}_1$ related to the relation

$$\boldsymbol{v}_1 = \boldsymbol{y}_1 - \boldsymbol{C}\,\boldsymbol{x}_1 = \boldsymbol{y}_1 - \boldsymbol{C}\,(\boldsymbol{A}\,\boldsymbol{x}_0 + \boldsymbol{B}\,\boldsymbol{u}_0 + \boldsymbol{G}\,\boldsymbol{w}_0) \tag{18.49}$$

The optimized variables being denoted as $\boldsymbol{x}_0^*$ and $\boldsymbol{w}_0^*$, this yields the predicted state

$$\hat{\boldsymbol{x}}_1 = \boldsymbol{A}\,\boldsymbol{x}_0^* + \boldsymbol{B}\,\boldsymbol{u}_0 + \boldsymbol{G}\boldsymbol{w}_0^* \tag{18.50}$$

A drawback of full state estimation is that the size of the optimization problem grows as time increases, which would very likely induce a failure in the optimization or pose problems in real-time control. The logical solution to this increasing size is to set the problem according to the moving-horizon approach.

Now consider the problem of moving state estimation. The criterion (18.47) is split into two parts (Rao et al., 2001; Robertson et al., 1996)

$$J_k = J_{k-H} + \sum_{i=k-H}^{k-1} (\boldsymbol{v}_{i+1}^T \boldsymbol{R}^{-1} \boldsymbol{v}_{i+1} + \boldsymbol{w}_i^T \boldsymbol{Q}^{-1} \boldsymbol{w}_i) = J_{k-H} + J^{mhe} \tag{18.51}$$

Clearly, the second part J^{mhe} of the criterion (18.51) depends on the state $\boldsymbol{x}_{k-H}$ and on the sequence of noises $\{\boldsymbol{w}_{k-H}, \ldots, \boldsymbol{w}_{k-1}\}$. The parallel with dynamic optimization (section 14.5) appears in this splitting as J_{k-H} itself must be optimized. If $k \leq H$, the problem is a full state estimation problem. Assume that $k > H$ and set the optimized criterion

$$J_{k-H}^* = \min_{\boldsymbol{x}_0, \boldsymbol{w}_0, \ldots, \boldsymbol{w}_{k-H-1}} J_{k-H} \tag{18.52}$$

so that, in the full optimized criterion becomes

$$\begin{aligned} J_k^* &= \min_{\boldsymbol{x}_0, \boldsymbol{w}_0, \ldots, \boldsymbol{w}_{k-1}} J_k \\ &= \min_{\boldsymbol{z}, \boldsymbol{w}_{k-H}, \ldots, \boldsymbol{w}_{k-1}} \left[\sum_{i=k-H}^{k-1} (\boldsymbol{v}_{i+1}^T \boldsymbol{R}^{-1} \boldsymbol{v}_{i+1} + \boldsymbol{w}_i^T \boldsymbol{Q}^{-1} \boldsymbol{w}_i) \right] + J_{k-H}^*(\boldsymbol{z}) \end{aligned} \tag{18.53}$$

where $\boldsymbol{z}$ is the arrival state $\boldsymbol{x}_{k-H}$ based on the optimized variables $\boldsymbol{x}_0^*$ and $\{\boldsymbol{w}_{k-H}^*, \ldots, \boldsymbol{w}_{k-H-1}^*\}$.

In actual practice, it is not possible to really minimize $J_{k-H}(z)$ when k becomes large as this would be again a full estimation problem. The solution

is to retain the previous values of the optimized criterion J_k^* obtained by moving horizon estimation denoted by $J_k^{mhe}(\boldsymbol{z})$ along time k and to approximate $J_{k-H}(\boldsymbol{z})$ as

$$J_{k-H}(\boldsymbol{z}) \approx (\boldsymbol{z} - \hat{\boldsymbol{x}}_{k-H}^{mhe})^T \, \boldsymbol{\Pi}_{k-H}^{-1} \, (\boldsymbol{z} - \hat{\boldsymbol{x}}_{k-H}^{mhe}) + J_{k-H}^{mhe}(\boldsymbol{z}) \tag{18.54}$$

where $\hat{\boldsymbol{x}}_{k-H}^{mhe}$ is the state estimated by moving horizon estimation at time $(k - H)$. Under these hypotheses, the criterion (18.51) becomes

$$J_k = \sum_{i=k-H}^{k-1} (\boldsymbol{v}_{i+1}^T \, \boldsymbol{R}^{-1} \, \boldsymbol{v}_{i+1} + \boldsymbol{w}_i^T \, \boldsymbol{Q}^{-1} \, \boldsymbol{w}_i) + (\boldsymbol{z} - \hat{\boldsymbol{x}}_{k-H}^{mhe})^T \, \boldsymbol{\Pi}_{k-H}^{-1} \, (\boldsymbol{z} - \hat{\boldsymbol{x}}_{k-H}^{mhe}) + J_{k-H}^{mhe}(\boldsymbol{z}) \tag{18.55}$$

The discrete Riccati equation (11.92) used for the covariance matrix of the Kalman filter is called to update $\boldsymbol{\Pi}_k$

$$\boldsymbol{\Pi}_k = \boldsymbol{A} \, \boldsymbol{\Pi}_{k-1} \, \boldsymbol{A}^T + \boldsymbol{G} \, \boldsymbol{Q} \, \boldsymbol{G}^T - \boldsymbol{A} \, \boldsymbol{\Pi}_{k-1} \, \boldsymbol{C}^T \, [\boldsymbol{C} \, \boldsymbol{\Pi}_{k-1} \, \boldsymbol{C}^T + \boldsymbol{R}]^{-1} \boldsymbol{C} \, \boldsymbol{\Pi}_{k-1}^T \, \boldsymbol{A}^T$$
with: $\boldsymbol{\Pi}_0$ given
(18.56)

Algorithm of the moving state estimation:

1/ Initialization of the states, of the matrices of the criterion.

2/ Get the measurements $\boldsymbol{y}(k)$. Store the previous manipulated inputs and measured outputs.

3/ Compare the current time index k to the value of the moving horizon H. If $k \leq H$, perform full state estimation. If $k > H$, perform moving state estimation.

4/ State estimation:

- 4a/ Call the initial parameters for nonlinear optimization.
- 4b/ Perform the nonlinear optimization.
- 4c/ Calculate the new estimed states $\hat{\boldsymbol{x}}$.
- 4d/ Update the covariance matrix $\boldsymbol{\pi}$.
- 4e/ Store the estimated states (and the optimal value of the criterion).
- 4f/ Go back to stage 2.

Example 18.1: **Moving horizon state estimation for a biological process**

The studied biological process is a fed-batch bioreactor where three states are considered: biomass (X), substrate (S) and product (P). Only the substrate is measured. The biomass cannot be measured on-line without noticeable delay and the product is in general only measurable with a rather important delay.

The model of this biological process is formulated under nonlinear continuous state-space form and comes from mass balances and equations of biological

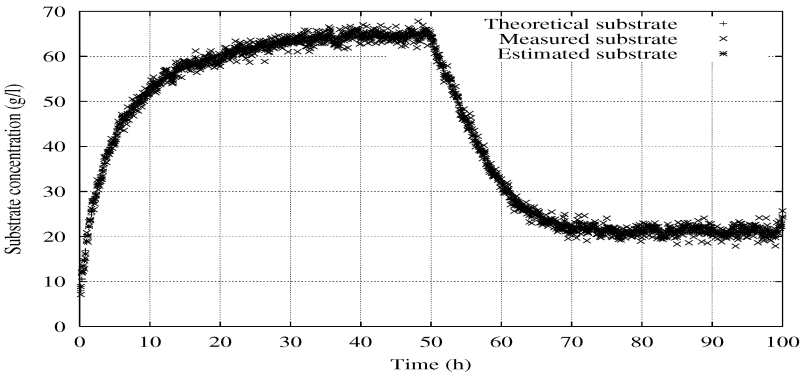


Figure 18.3: Moving state estimation of substrate

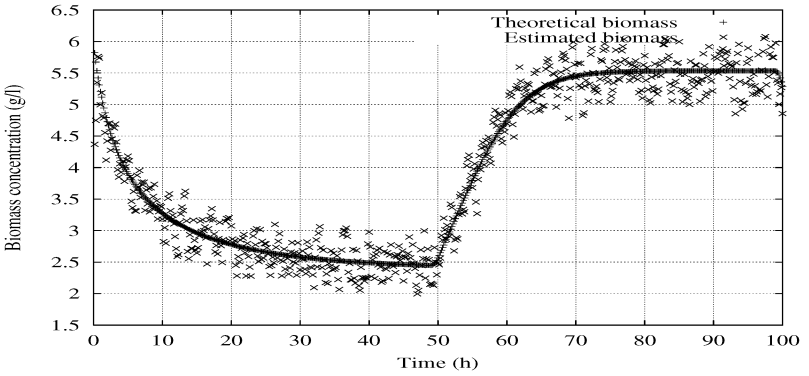


Figure 18.4: Moving state estimation of biomass

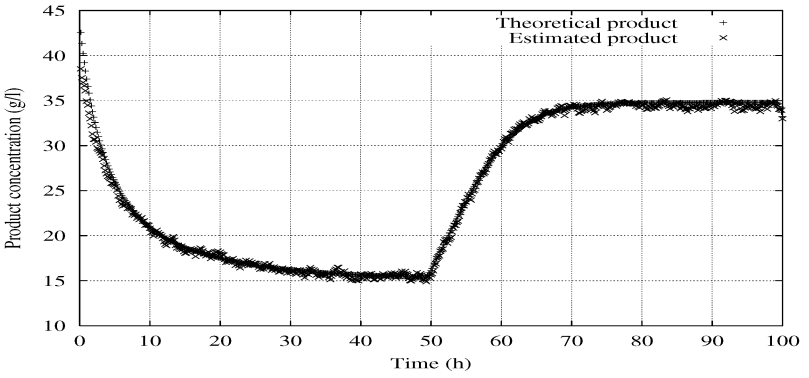


Figure 18.5: Moving state estimation of product

18.5 Conclusion

State and parameter observers are capable of providing a significant help to operators for process monitoring and diagnostic. In the case of linear or non-

A. Isidori. *Nonlinear Control Systems.* Springer-Verlag, New York, 3rd edition, 1995.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D, J. Basic Eng.*, 82:35–45, 1960.

R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME Ser. D, J. Basic Eng.*, 83:95–108, 1961.

M. Kano, K. Miyazaki, S. Hasebe, and I. Hashimoto. Inferential control system of distillation compositions using dynamic partial least squares regression. *J. Proc. Cont.*, 10:157–166, 2000.

M.H. Kaspar and W.H. Ray. Chemometric methods for process monitoring and high-performance controller design. *AIChE J.*, 38(10):1593–1608, 1992.

A.J. Krener and A. Isidori. Linearization by output injection and nonlinear observers. *Systems and Control Letters*, 3(1):47–52, 1983.

J.V. Kresta, J.F. MacGregor, and T.E. Marlin. Multivariate statistical monitoring of process operating performance. *Can. J. Chem. Eng.*, 69(2):35–47, 1991.

W. H. Kwon, P. S. Kim, and P. G. Park. A receding horizon Kalman FIR filter for discrete time-invariant systems. *IEEE TAC*, 44(9):1787–1791, 1999.

S. Lakshminarayanan, S.L. Shah, and K. Nandakumar. Modeling and control of multivariable process: dynamic PLS approach. *AIChE J.*, 43(9):2307–2322, 1997.

R. Li and J.H. Olson. Fault detection and diagnosis in a closed-loop nonlinear distillation process: Application of extended Kalman filters. *Ind. Eng. Chem. Res.*, 30:898–908, 1991.

L. Ljung. Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems. *IEEE Trans. Automat. Control*, AC-24(1):36–50, 1979.

D.G. Luenberger. Observers for multivariable systems. *IEEE Trans. Automat. Control*, AC-11(2):190–197, 1966.

D.G. Luenberger. An introduction to observers. *IEEE Transactions on Automatic Control*, AC-16(6):596–602, 1971.

D.G. Luenberger. *Introduction to Dynamic Systems. Theory, Models and Applications.* Wiley, New York, 1979.

T. Mejdell and S. Skogestad. Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression. *Ind. Eng. Chem. Res.*, 30:2543–2555, 1991a.

T. Mejdell and S. Skogestad. Composition estimator in a pilot-plant distillation column using multiple temperatures. *Ind. Eng. Chem. Res.*, 30:2555–2564, 1991b.

H. Michalska and D. Q. Mayne. Moving horizon observers and observer-based control. *IEEE TAC*, 40:995–1006, 1995.

E.A. Misawa and J.K. Hedrick. Nonlinear observers - A state-of-the-art survey. *Transactions of the ASME*, 111:344–352, 1989.

M. Otto. *Chemometrics*. Wiley-VCH, Weinheim, 1999.

C. V. Rao, J. B. Rawlings, and J. H. Lee. Constrained linear state estimation - a moving horizon approach. *Automatica*, 37:1619–1628, 2001.

A.C. Rencher. *Methods of Multivariate Analysis*. Wiley, New York, 1995.

D. G. Robertson, J. H. Lee, and J. B. Rawlings. A moving horizon based-approach for least-squares estimation. *AIChE J.*, 42(8):2209–2224, 1996.

M.G. Safonov and M. Athans. Robustness and computational aspects of nonlinear stochastic estimators and regulators. *IEEE Trans. Automat. Control*, AC-23(4):717–725, 1978.

K. Schittkowski. NLPQL: A Fortran subroutine solving constrained nonlinear programming problems. *Ann. Oper. Res.*, 5:485–500, 1985.

J. E. Slotine, J. K. Hedrick, and E. A. Misawa. On sliding observers for nonlinear systems. *Journal of Dynamics Systems, Measurements and Control*, 109:245–252, 1987.

T. Söderström and P. Stoica. *System Identification*. Prentice Hall, New York, 1989.

M. Tenenhaus. *La Régression PLS. Théorie et Pratique*. Technip, Paris, 1998.

J. Trygg and S. Wold. Orthogonalized projections to latent structures, O-PLS. *J. Chemometrics*, 16:119–128, 2002.

G. B. Wang, S. S. Peng, and H. P. Huang. A sliding observer for nonlinear process control. *Chem. Eng. Sci.*, 52:787–805, 1997.

D.I. Wilson, M. Agarwal, and D.W. T. Rippin. Experiences implementing the extended Kalman filter on an industrial batch reactor. *Comp. Chem. Engng.*, 22(11):1653–1672, 1998.

S. Wold, A. Ruhe, H. Wold, and W.J. Dunn III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, 5(3):735–743, 1984.

S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chem. Intell. Lab. Sys.*, 2:37–52, 1987.

S. Wold, J. Trygg, A. Berglund, and H. Antti. Some recent developments in PLS modeling. *J.Chemometrics*, 2001.

G. Zimmer. State observation by on-line minimization. *Int. Journal of Control*, 60:595–606, 1994.

Part VI

Applications to Processes

Chapter 19

Nonlinear Control of Reactors with State Estimation

19.1 Introduction

Many linear and nonlinear control methods exist. Very often, the choice of a particular method depends on the theoretical knowledge of the person responsible for the development of the control system. This choice also depends on the characteristics of the studied system. Linear control such as pole-placement (Sect. 13.1), internal model control (Sect. 13.2) and generalized predictive control (Sect. 15.6), have been previously studied using an identified linear model (Sect. 12.7.2) of the chemical reactor described in Sect. 19.2.

Exothermic chemical reactors, especially polymerization reactors, present several stationary points and must frequently be controlled around an unstable point. Their characteristics are highly nonlinear, particularly for batch reactors.

Biological reactors are most often operated in batch or semi-batch mode, especially to avoid contamination problems. Their behaviour can only be described by complex kinetic models that are able to represent them in a large operating domain.

For the reasons previously cited, nonlinear geometric control is well adapted for such systems. It allows us to fully exploit the knowledge that the engineer has about the process. However, it necessitates us knowing the system states, and a state observer needs to be coupled.

19.2 Chemical Reactor

Examples of nonlinear geometric control can be found for an exothermic reactor (Limqueco and Kantor, 1990), for control of batch reactors (Kravaris

et al., 1989; Soroush and Kravaris, 1992), with feedforward-feedback (Daoutidis et al., 1990), for a multivariable continuous polymerization reactor (Soroush and Kravaris, 1993, 1994).

The example retained here is classical with regard to the mass and heat balances describing the dynamic behaviour of a chemical reactor. It concerns a continuous reactor in which a single reaction occurs: A $\longrightarrow$ B . The nonlinear state-space model is composed of three differential equations describing the evolution of the three states (C_A concentration in A, T reactor temperature, T_j jacket temperature), among which only one is measured (the reactor temperature).

19.2.1 Model of the Chemical Reactor

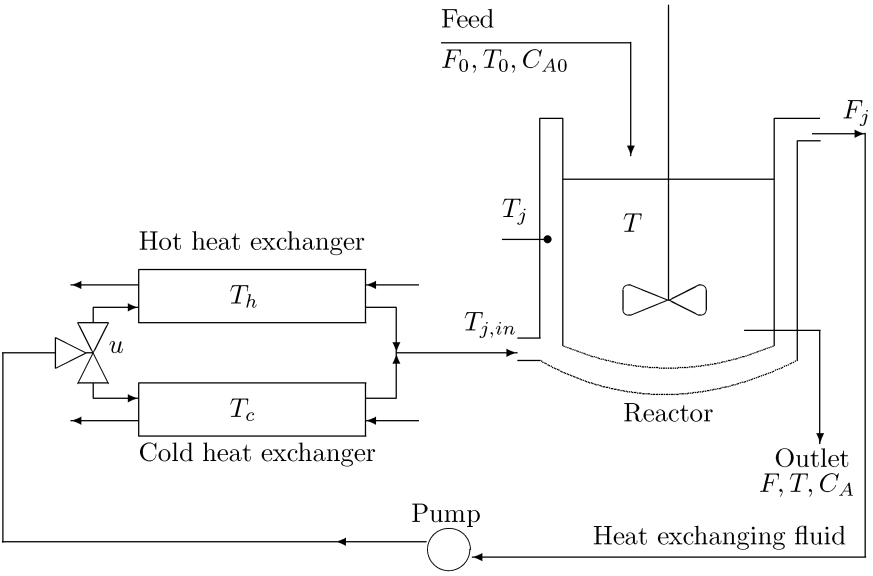


Figure 19.1: Scheme of the chemical reactor with jacket and heat exchanger system acting on the temperature entering the jacket

The continuous chemical reactor (inlet with subscript "0", outlet having the properties of the reactor contents) is assumed to be perfectly stirred. The n-order reaction $A \longrightarrow B$ releases a heat of reaction ΔH. The reactor (Fig. 19.1) is surrounded by a jacket (subscript "j") of constant volume V_j crossed by the heat-conducting fluid circulating at a constant flow rate F_j with a variable inlet temperature $T_{j,in}$. Corresponding to a lab pilot reactor, the inlet jacket stream results from the mixing of two streams, one running through a cold heat exchanger (temperature T_c), the other one running through a hot heat exchanger (temperature T_h). The proportion of fluid through these exchangers depends on the position u of a three-way valve. Thus, the inlet jacket tempe-

rature is grossly a weighted function of both heat exchanger temperatures. A proportional feedback controller regulates the liquid level in the reactor around a volume set point V_c (Luyben, 1990). The features of this reactor for the heat exchanger design are very close to the pilot reactor used by Gentric et al. (1999).

The equations ruling this reactor are the following:

- Inlet temperature in the jacket

$$T_{j,in} = u\,T_h + (1-u)\,T_c \tag{19.1}$$

- Mass balance of the reactor

$$dV/dt = F_0 - F \tag{19.2}$$

- Balance in component A

$$d(C_A)/dt = F_0\,(C_{A0} - C_A)/V - k\,C_A^n \tag{19.3}$$

- Energy balance for the reactor

$$dT/dt = F_0/V\,(T_0 - T) - \Delta H\,k\,C_A^n/(\rho\,C_p) - U\,A\,(T - T_j)/(V\,\rho\,C_p) \tag{19.4}$$

- Energy balance for the jacket

$$d(T_j)/dt = F_j/V_j\,(T_{j,in} - T_j) + U\,A\,(T - T_j)/(V_j\,\rho_j\,C_{pj}) \tag{19.5}$$

- Proportional level control in the reactor

$$F - F^s = \delta F = -K_v\,(V_c - V) \tag{19.6}$$

- Kinetic constant and activation energy

$$k = k_0\,\exp(-E/RT) \tag{19.7}$$

To simplify, we assume that the reaction order n is equal to 1.

The reactor characteristics are given in Table 19.1.

19.2.2 Control Problem Setting

The control aim is to make the reactor contents follow a temperature profile.

From the previous differential equations, the state vector is

$$\boldsymbol{x} = [\,x_1 = C_A\ ;\ x_2 = T\ ;\ x_3 = T_j\ ;\ x_4 = V\,]^T \tag{19.8}$$

The state-space model is

$$\begin{aligned}
\dot{x}_1 &= \frac{F_0\,(C_{A0} - x_1)}{x_4} - k_0 \exp(-\frac{E}{Rx_2})\,x_1 \\
\dot{x}_2 &= \frac{F_0\,(T_0 - x_2)}{x_4} - \frac{\Delta H\,k_0 \exp(-\frac{E}{Rx_2})\,x_1}{\rho\,C_p} - \frac{U\,A\,(x_2 - x_3)}{x_4\,\rho\,C_p} \\
\dot{x_3} &= \frac{F_j\,(T_{j,in} - x_3)}{V_j} + \frac{U\,A\,(x_2 - x_3)}{V_j\,\rho_j\,C_{pj}} \\
\dot{x_4} &= F_0 - F
\end{aligned} \tag{19.9}$$

Table 19.1: Initial variables and main parameters of the CSTR

Flow rate of the feed	$F_0 = 3 \times 10^{-4}$ m^3.s^{-1}
Concentration of reactant A in the feed	$C_{A0} = 3900$ mol.m^{-3}
Temperature of the feed	$T_0 = 295$ K
Volume of reactor	$V = 1.5$ m^3
Kinetic constant	$k_0 = 2 \times 10^7$.s^{-1}
Activation energy	$E = 7 \times 10^4$ J.mol^{-1}
Heat of reaction	$\Delta H = -7 \times 10^4$ J.mol^{-1}
Density of reactor contents	$\rho = 1000$ kg.m^{-3}
Heat capacity of reactor contents	$C_p = 3000$ J.kg^{-1}.K^{-1}
Temperature of the cold heat exchanger	$T_c = 280$ K
Temperature of hot heat exchanger	$T_h = 360$ K
Flow rate of the heat-conducting fluid	$F_j = 5 \times 10^{-2}$ m^3.s^{-1}
Volume of jacket	$V_j = 0.1$ m^3
Heat transfer coefficient between the jacket and the reactor contents	$U = 900$ W.m^{-2}.K^{-1}
Heat-exchange area	$A = 20$ m^2
Density of the heat-conducting fluid	$\rho_j = 1000$ kg.m^{-3}
Heat capacity of the heat-conducting fluid	$C_{pj} = 4200$ J.kg^{-1}.K^{-1}
Proportional gain of the level controller	$K_v = 0.05$ s^{-1}

$$y = x_2 \tag{19.10}$$

To maintain the reactor around its equilibrium state, a nonlinear control based on differential geometry is used. It necessitates us knowing the states, but two of these are not measured. Thus, an observer or state estimator must be implemented. In the present case, we will use the extended Kalman filter (Watanabe, 1992) as the reactor dynamic model is nonlinear. If this model were linear, the linear Kalman filter would have been applied. On the other hand, the knowledge of the concentration in the reactor based only on the measurement of the temperature can be important information for the operators running the process. Notice that a minimum model of the process is necessary, and that the better the model, the better will be the estimation for a given measurement quality. In the same arena of ideas, the estimation of more complex variables concerning polymerization reactions such as the polydispersity and the molecular weight distribution moments is possible (Gentric et al., 1999).

19.2.3 Control Law

By replacing the inlet jacket temperature $T_{j,in}$ by its expression with respect to the position u of the valve, the dynamic model (19.9) of the chemical reactor is expressed as a nonlinear system, single-input single-output, and affine with

respect to the control input, as

$$\begin{cases} \dot{x} = & f(x) + g(x)u \\ y = & h(x) \end{cases} \tag{19.11}$$

with the following vector fields

$$f(x) = \begin{bmatrix} \dfrac{F_0\,(C_{A0} - x_1)}{x_4} - k_0 \exp(-\dfrac{E}{Rx_2})\,x_1 \\ \dfrac{F_0\,(T_0 - x_2)}{x_4} - \dfrac{\Delta H\,k_0 \exp(-\dfrac{E}{Rx_2})\,x_1}{\rho\,C_p} - \dfrac{U\,A\,(x_2 - x_3)}{x_4\,\rho\,C_p} \\ \dfrac{F_j\,(T_f - x_3)}{V_j} + \dfrac{U\,A\,(x_2 - x_3)}{V_j\,\rho_j\,C_{pj}} \\ F_0 - F \end{bmatrix} \tag{19.12}$$

$$g(x) = \begin{bmatrix} 0 \\ 0 \\ \dfrac{F_j\,(T_c - T_f)}{V_j} \\ 0 \end{bmatrix} \tag{19.13}$$

and the scalar output

$$h(x) = x_2 \tag{19.14}$$

First, the relative degree of this system is to be determined. For that, the following Lie derivatives are calculated:

– The Lie derivative of $h(x)$ in the direction of the vector field f

$$\begin{aligned} L_f h(x) &= \frac{\partial h}{\partial x}\,f(x) = f_2(x) \\ &= \frac{F_0}{x_4}\,(T_0 - x_2) - \frac{\Delta H}{\rho\,C_p}\,k_0 \exp(-E/Rx_2)\,x_1 - \frac{U\,A}{x_4\,\rho\,C_p}\,(x_2 - x_3) \end{aligned} \tag{19.15}$$

– The Lie derivative of $h(x)$ in the direction of the vector field g

$$L_g L_f^0 h(x) = L_g h(x) = 0 \tag{19.16}$$

which implies that the relative degree is larger than 1.

– The Lie derivative of $L_f h(x)$ in the direction of the vectors field g

$$L_g L_f h(x) = \frac{\partial L_f h(x)}{\partial x}\,g(x) = \frac{U\,A}{x_4\,\rho\,C_p}\,F_j/V_j\,(T_c - T_f) \tag{19.17}$$

This derivative is always different from zero, as the characteristic temperatures of the cold and hot heat exchangers are different. The relative degree of the output $h(x) = x_2$ of the nonlinear system (19.9) is thus equal to 2, i.e. it is necessary to twice differentiate the temperature to make the input appear linearly.

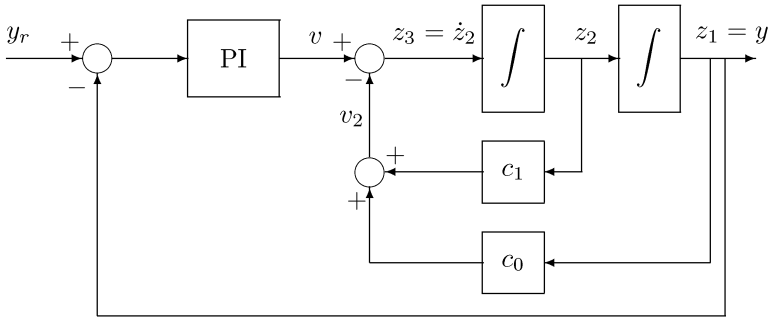


Figure 19.2: Nonlinear geometric control of the chemical reactor with pole-placement

In order to eliminate the stationary error, a PI controller is introduced so that the external input is equal to

$$v(t) = K_c \left[(y_r(t) - y(t)) + \frac{1}{\tau_I} \int_0^t (y_r(\tau) - y(\tau)) d\tau \right] \tag{19.18}$$

where y_r is the set point and y the controlled output. The complete system is represented by the block diagram in Fig. 19.2.

The general form of the state feedback control law that linearizes the input-output behaviour of the closed-loop system (Isidori, 1995) results

$$u = \frac{v - c_0 h(x) - c_1 L_f h(x) - L_f^2 h(x)}{L_g L_f h(x)} \tag{19.19}$$

where c_0, c_1 are constant scalar parameters. The Lie derivative $L_f^2 h(x)$ is only presented in an abbreviated form, which can be easily expressed

$$L_f^2 h(x) = \sum_i \frac{\partial f_2}{\partial x_i} f_i(x) \tag{19.20}$$

Of course, it is necessary to modify Eq. (19.19) so as to take into account the reference trajectory as in Eq. (17.133).

Then, it is sufficient to determine the parameters c_0, c_1, K_c, τ_I so that the closed-loop transfer function equal to

$$\frac{Y(s)}{Y_r(s)} = \frac{K_c(s + 1/\tau_I)}{s^3 + c_1 s^2 + (c_0 + K_c)s + K_c/\tau_I} \tag{19.21}$$

presents the desired characteristics, and thus that the poles of the following characteristic equation

$$s^3 + c_1 s^2 + (c_0 + K_c)s + K_c/\tau_I = 0 \tag{19.22}$$

are correctly located in the complex plane; this corresponds to a pole-placement. For this, it is possible to use an optimal continuous polynomial in the sense of the ITAE criterion (Table 4.1).

19.2.4 State Estimation

Only the temperature and volume of the reactor contents are assumed to be measured, which is often the case in practice. The extended Kalman filter was used to estimate the three states: concentration C_A, reactor temperature T and jacket temperature T_j. To the theoretical reactor temperature given by the simulation model was added Gaussian white noise of standard deviation 0.1 to simulate the measurement errors.

The algorithm of the extended Kalman filter (Sect. 18.4.3) consists of two stages: a stage of state and error covariance matrix prediction and a stage of correction of the predicted variables. In the prediction stage, the system of differential equations describing the derivatives of the states and the derivatives of the elements of the covariance matrix is integrated. The correction stage modifies the predicted values by adding a term such as

$$K(t)\,[y_{\text{measured}}(t) - y_{\text{predicted}}(t)] \tag{19.23}$$

where $K(t)$ is the Kalman gain matrix calculated from the predicted values and $y(t)$ a measured output of the process.

By adopting the general notations of Eqs. (18.16) and (18.17), the vector h defining the output is equal to

$$h = [0\ ;\ x_2\ ;\ 0\ ;\ 0] \tag{19.24}$$

We consider that the level is perfectly known and that the reactor volume is not estimated, so that $\partial/\partial x_4 = 0$. We calculate the Jacobian matrix (18.21). To facilitate the writing, we use the Kronecker symbol: $\delta_{ij} = 1$ if $i = j \quad ; \delta_{ij} = 0$ if $i \neq j$.

The partial derivatives ($i = 1, 2, 3, 4$) are then equal to

$$\begin{aligned}
F_{1i} = \frac{\partial(f_1 + g_1\,u)}{\partial x_i} &= -\frac{F_0}{x_4}\,\delta_{1i} - k\,\delta_{1i} - \frac{E}{Rx_2^2}\,\delta_{2i}\,k\,x_1 \\
F_{2i} = \frac{\partial(f_2 + g_2\,u)}{\partial x_i} &= -\frac{F_0}{x_4}\,\delta_{2i} - \frac{\Delta H}{\rho\,C_p}\,k\,\delta_{1i} \\
&\quad -\frac{\Delta H}{\rho\,C_p}\,\frac{E}{Rx_2^2}\,\delta_{2i}\,k\,x_1 - \frac{U\,A}{x_4\,\rho\,C_p}\,(\delta_{2i} - \delta_{3i}) \\
F_{3i} = \frac{\partial(f_3 + g_3\,u)}{\partial x_i} &= -\frac{F_j}{V_j}\,\delta_{3i} + \frac{U\,A}{V_j\,\rho_j\,C_{pj}}\,(\delta_{2i} - \delta_{3i}) \\
F_{4i} = \frac{\partial(f_4 + g_4\,u)}{\partial x_i} &= 0
\end{aligned} \tag{19.25}$$

a) Realization of the propagation stage:
The error covariance matrix $\boldsymbol{P}$ is initialized as a scalar matrix of elements

$$P_{ii} = (\hat{x}_i/20)^2; \quad P_{ij} = 0 \text{ if: } i \neq j \tag{19.26}$$

while the state noise covariance matrix $\boldsymbol{Q}$ is also scalar and equal to

$$Q_{ii} = 10^{-4}; \quad Q_{ij} = 0 \text{ if: } i \neq j \tag{19.27}$$

Equation (18.19) is the system constituted by the first four differential equations of model (19.9) after having replaced x by $\hat{x}$.
Equation (18.20) is a system of 4×4 differential equations obtained after having replaced $\boldsymbol{F}$, the Jacobian matrix calculated by Eq. (19.25) and $\boldsymbol{Q}$ previously initialized.

Thus, we integrate numerically this system of (4+9) differential equations on a sampling period. We then obtain $\hat{\boldsymbol{x}}_k(-)$ and $\boldsymbol{P}_k(-)$.
b) Realization of the update stage:

The Jacobian matrix $\boldsymbol{H}$ of h, which is, in fact, a scalar function, is to be calculated. Thus, $\boldsymbol{H}$ is a vector and its only nonzero element is $H_{12} = 1$.

The noise $\boldsymbol{R}_k$ in the measurement is a diagonal matrix. In fact, only the temperature is measured, thus $\boldsymbol{R}_k$ has dimension 1: $R_{22} = \sigma_2^2 = 10^{-2}$.

The Kalman gain matrix $\boldsymbol{K}_k$ is calculated from Eq. (18.25), then the state estimations from Eq. (18.22)

$$\begin{array}{lcl} \hat{C}_A = \hat{x}_1(+) & = & \hat{x}_1(-) + K_{12}\left[x_{2,mes} - \hat{x}_2(-)\right] \\ \hat{T} = \hat{x}_2(+) & = & \hat{x}_2(-) + K_{22}\left[x_{2,mes} - \hat{x}_2(-)\right] \\ \hat{T}_j = \hat{x}_3(+) & = & \hat{x}_3(-) + K_{32}\left[x_{2,mes} - \hat{x}_2(-)\right] \\ \hat{V}_j = V_j & & \end{array} \tag{19.28}$$

We also update the covariance matrix $\boldsymbol{P}_k(-)$ by means of Eq. (18.23) to obtain $\boldsymbol{P}_k(+)$.

Then we can restart at stage a).

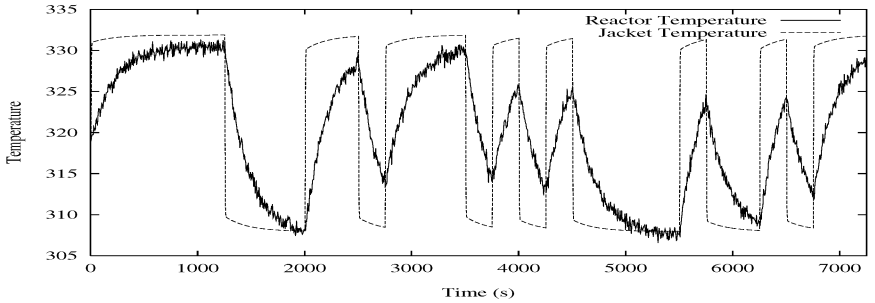


Figure 19.3: Open-loop response to a pseudo-random binary sequence: temperatures of the reactor and the jacket

19.2.5 Simulation Results

Two types of simulation are realized: in the absence and in the presence of chemical reaction. In the first case, only the phenomena of pure heat transfer occur. When the chemical reaction intervenes with its heat of reaction, the temperature control is influenced. For a batch reactor, the reactants are mixed only when the desired temperature is obtained and, at that instant, an important heat of reaction is developed in the reactor where a thermal runaway can

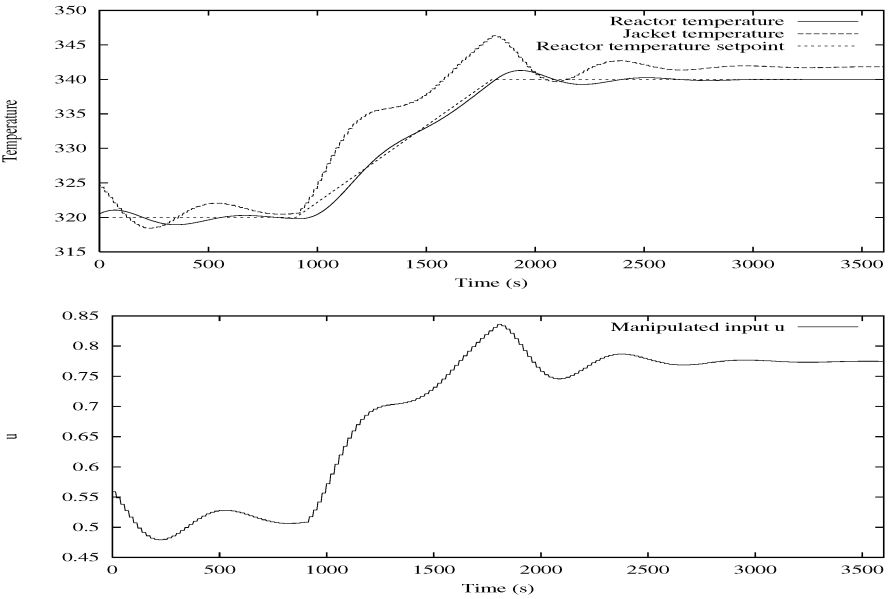


Figure 19.4: Nonlinear geometric control in the absence of chemical reaction. Top: temperatures of the reactor and of the jacket and temperature set point. Bottom: valve position

occur if the temperature increase is not controlled. For a continuous reactor, in nominal operation, the heat of reaction is a well mastered normal phenomenon.

Here, the characteristics of a continuous reactor have been modified to resemble those of a batch reactor: at the initial time, the reactor is assumed not to be at steady state and there is no chemical reaction. It can be compared to a continuous chemical reactor at the startup. After heating to the desired temperature, the chemical reaction is operated and the heat of reaction thus evolved must be controlled. In this case, in terms of control, this heat of reaction is a disturbance. For a continuous reactor operating normally, the heat of reaction would be identified with the rest of the heat transfer phenomena.

A first open-loop simulation in the absence of chemical reaction in response to a pseudo-random binary sequence applied to the valve position allows us to evaluate the reactor time constant related to the heat transfer essentially between the jacket and the reactor contents. The behaviour (Fig. 19.3) is close to that of a first-order linear system of time constant around $\tau = 250\,\mathrm{s}$.

The closed-loop simulation was realized in the following conditions: the sampling period is chosen to be equal to 5 s. The control input u varies in the interval $[0, 1]$, and is modified only every 20 s. The parameters of the control law are chosen so that the closed-loop linearized system is optimal for the ITAE criterion: $K_c = 6.8 \times 10^{-5}, \tau_I = 4.28, c_0 = 0.0013, c_1 = 0.044$. The integral term of PI control is only active when the deviation between set point

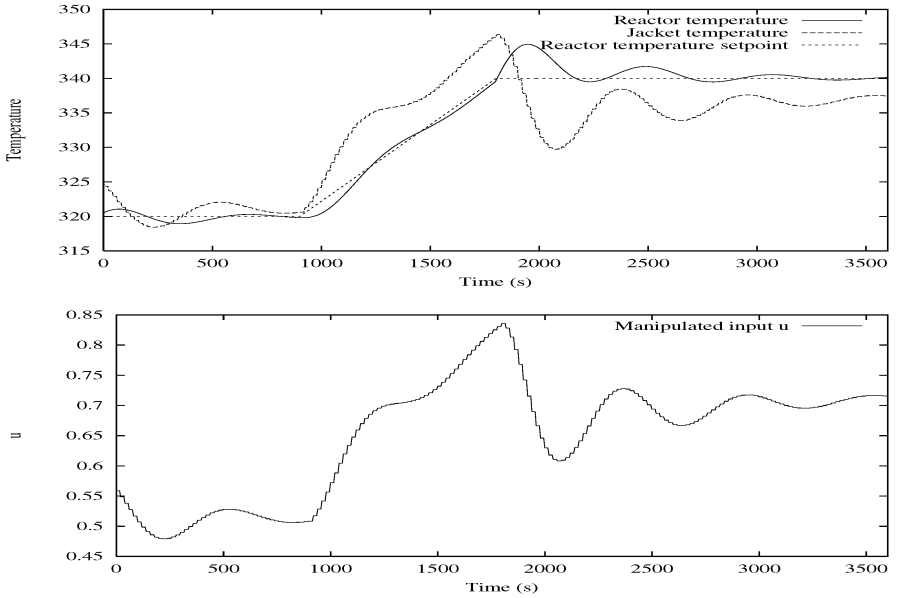


Figure 19.5: Nonlinear geometric control with chemical reaction starting at $t = 1800\,\text{s}$. Top: temperatures of the reactor and of the jacket and temperature set point. Bottom: valve position

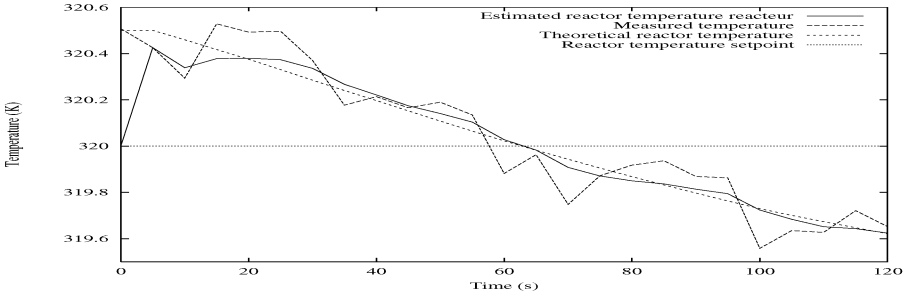


Figure 19.6: Nonlinear geometric control with chemical reaction starting at $t = 1800\,\text{s}$: theoretical, measured and estimated temperatures of the reactor

and measurement is lower than 2 K, in order to avoid the windup of the PI controller. The set point is always filtered by a second-order filter to avoid too brutal variations of the control input. Note that this reactor is not considered at steady-state at $t = 0$ and that in all closed-loop figures, the chemical reaction starts only at $t = 1800\,\text{s}$.

The initial theoretical states and the estimated initial states are given in Table 19.2. In this case, we do not consider possible changes in the feed flow rate and reactor volume.

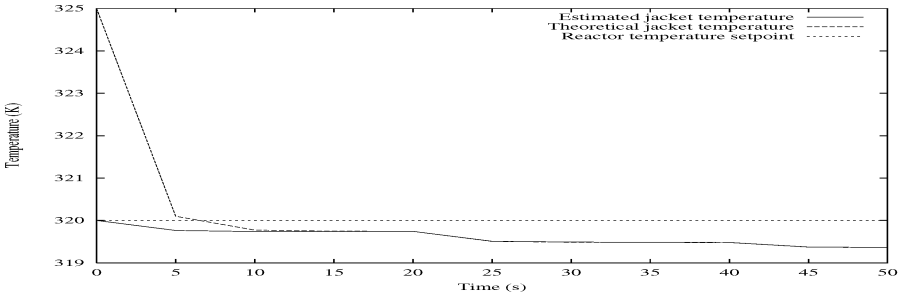


Figure 19.7: Nonlinear geometric control with chemical reaction starting at $t = 1800\,\text{s}$: theoretical and estimated jacket temperatures; temperature set point

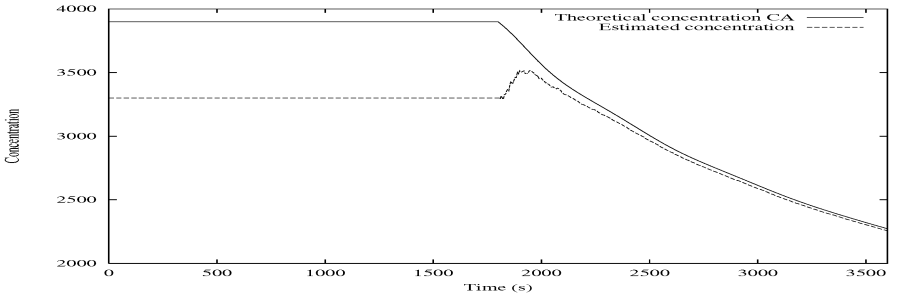


Figure 19.8: Nonlinear geometric control with chemical reaction starting at $t = 1800\,\text{s}$: theoretical and estimated concentrations

A first closed-loop simulation is performed in the absence of chemical reaction (Fig. 19.4). The desired profile is followed without difficulty and the corresponding control input varies slowly.

In the second closed-loop simulation, the reaction starts at $t = 1800\,\text{s}$, when the desired temperature level is reached. At this time, the highly exothermic reaction induces a noticeable temperature overshoot (Fig. 19.5). The return towards the stationary is realized with a few oscillations. It would be possible

Table 19.2: Initial theoretical states and initial estimated states of the CSTR

Initial state	Theoretical	Estimated
Concentration C_A	3900 mol/m^3	3300 mol/m^3
Reactor contents temperature T	320.5 K	320.0 K
Jacket temperature T_j	325.0 K	320.0 K
Reactor volume V	1.5 m^3	1.5 m^3

to limit the oscillations by modifying the controller parameters with regard to those obtained from the ITAE criterion. The control input still varies relatively slowly in spite of the disturbance provoked by the exothermic reaction.

Now consider the performances of the nonlinear observer. The continuous reactor is studied in a start stage and is not initially at equilibrium. The temperatures and concentrations thus tend towards this stationary state. We assume that only the initial measured temperature of the reactor contents is known with a small error. On the contrary, the large initial estimation errors concern both unmeasured states: the jacket temperature and the concentration (10 K for jacket temperature, 20% for the concentration). As this is only simulation, the nonlinear model of the reactor represents the theoretical plant and it is possible to compare a theoretical variable and the corresponding estimated variable. The estimation of the temperature (Fig. 19.6), which is also measured, shows that the observer convergence is fast and that the deviation between the measured temperature of the reactor and its estimated temperature is, of course, small, as the initial error is small and the correction uses the measured temperature of the reactor. The estimated jacket temperature also converges very rapidly towards the theoretical temperature (Fig. 19.7); this is explained as the reactor temperature (measured and noisy) and the jacket temperature are closely linked by the energy balance. The unmeasured and estimated concentration (Fig. 19.7) first tends rapidly, then more slowly towards the theoretical concentration (Fig. 19.8).

19.3 Biological Reactor

19.3.1 Introduction

Biotechnological processes are well known for their operation complexity and their difficult understanding related to the nature of living material. Knowledge models, which are, in general, non stationary, make use of many parameters and are strongly nonlinear. Moreover, the available measurements, in general, are not numerous and are of uncertain quality. However, the expansion of biotechnologies imposes development of high-performance control strategies that allow us to efficiently monitor these processes and improve their productivity.

To take into account the nonlinear and unsteady character of bioprocesses, a nonlinear control based on differential geometry and global linearization of the input-output behaviour seems appropriate (Isidori, 1989).

Nonlinear geometric control has been applied to biological reactors in particular by Hoo and Kantor (1986) and Pröll and Karim (1994).

The described application concerns an industrially important fermentation process constituted by a fed-batch reactor of baker's yeast production (Lucena et al., 1995, 2001). A nearly optimal production of biomass is obtained by prescribing to follow a low glucose concentration set point.

19.3.2 Dynamic Model of the Biological Reactor

The dynamic model of the baker's yeast production process in a fed-batch reactor (Fig. 19.9) is a physiological model of growth of *Saccharomyces cerevisiae* (Dantigny, 1989; Rajab, 1983). It relies on the existence of three limit physiological states of biomass (Fig. 19.10).

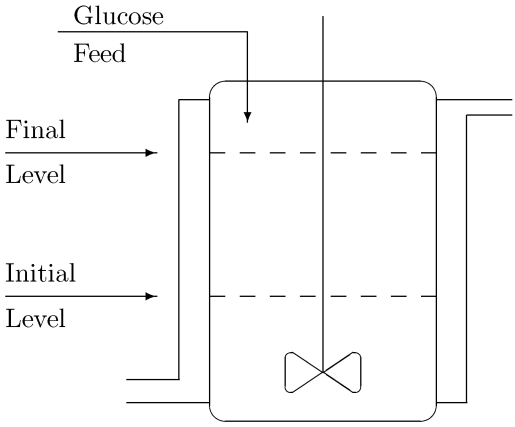


Figure 19.9: Fed-batch fermentor

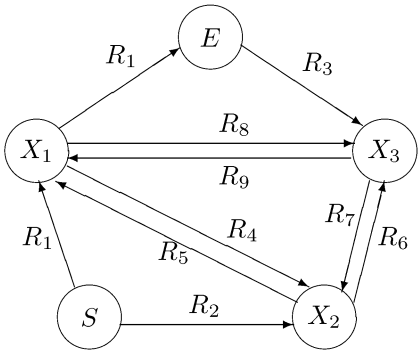


Figure 19.10: Reaction scheme (E: ethanol, S: substrate (glucose), X: biomass)

These physiological states X_1, X_2 and X_3 correspond respectively to the glucose fermentation, glucose respiration and ethanol respiration. The model kinetic expressions correspond to aerobic growth, in the absence of other limitations than that of substrate. The state vector x has dimension 7. It is formed by concentrations (g/l) of ethanol E, yeast X_1, X_2 and X_3, glucose S, acetate A, and the reactor liquid volume V (l).

The dynamic model of the reactor is represented by the following nonlinear

system, in the state space, affine with respect to the control input

$$\begin{cases} \dot{E} = & \frac{0.45}{0.14}R_1 - (1-\phi)\frac{R_3}{0.55} - \frac{E}{V}F_{in} \\ \dot{X}_1 = & R_1 + R_5 + R_9 - R_4 - R_8 - \frac{X_1}{V}F_{in} \\ \dot{X}_2 = & R_2 + R_4 + R_7 - R_5 - R_6 - \frac{X_2}{V}F_{in} \\ \dot{X}_3 = & R_3 + R_8 + R_6 - R_9 - R_7 - \frac{X_3}{V}F_{in} \\ \dot{S} = & -\frac{R_1}{0.14} - \frac{R_2}{0.5} - \frac{S}{V}F_{in} + \frac{S_{in}}{V}F_{in} \\ \dot{A} = & \frac{0.01}{0.14}R_1 - \phi\frac{R_3}{0.55} - \frac{A}{V}F_{in} \\ \dot{V} = & F_{in} \end{cases} \tag{19.29}$$

The control input is the glucose feed flow rate $u = F_{in}$, the output is the glucose concentration in the fermentor $y = S$.
R_1, R_2 and R_3 are biomass production rates defined by

$$\begin{aligned} R_1 = & \frac{0.6X_1S}{(0.5+S)(1+\frac{A}{0.4})} \\ R_2 = & \frac{0.29X_2S}{(0.04+S)} \\ R_3 = & \frac{0.25X_3}{(1+\frac{A}{0.2})}\left[\frac{E}{(0.02+E)} + \phi\frac{A}{(0.02+A)}\right] \end{aligned} \tag{19.30}$$

R_4, R_5, R_6, R_7, R_8 and R_9 are the transition rates from one state to another one defined by

$$\begin{aligned} R_4 &= 2\,X_1\,(1-\alpha) & R_5 &= 0.1\,X_2\,\alpha \\ R_6 &= \frac{0.4\,X_2\,E}{0.5+E} & R_7 &= \frac{2\,X_3\,S}{0.05+S} \\ R_8 &= \frac{0.2\,X_1\,E}{0.2+E}(1-\alpha) & R_9 &= \frac{0.5\,X_3\,S}{0.01+S} \end{aligned} \tag{19.31}$$

The rates R_1, R_2, ..., R_9 are expressed in g/(l.h), the dimensionless coefficients α and ϕ are defined as follows

$$\begin{aligned} &\alpha = \frac{S^3}{0.1^3+S^3} \\ &\begin{cases} \phi = 1 & \text{if } E \le 10^{-6}\ \text{g/l} \\ \phi = 0 & \text{if } E > 10^{-6}\text{g/l} \end{cases} \end{aligned} \tag{19.32}$$

Thus, the model is highly nonlinear.

19.3.3 Synthesis of the Nonlinear Control Law

The dynamic model (19.36) of the biological reactor is a SISO nonlinear analytic system, affine with respect to the control input, which can be written as

$$\begin{cases} \dot{x} = & f(x) + g(x)u \\ y = & h(x) \end{cases} \tag{19.33}$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} \qquad g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_n(x) \end{bmatrix} \tag{19.34}$$

with

$$x = [E \quad X_1 \quad X_2 \quad X_3 \quad S \quad A \quad V]^T \tag{19.35}$$

$$f(x) = \begin{bmatrix} \frac{0.45}{0.14}R_1 - (1-\phi)\frac{R_3}{0.55} \\ R_1 + R_5 + R_9 - R_4 - R_8 \\ R_2 + R_4 + R_7 - R_5 - R_6 \\ R_3 + R_8 + R_6 - R_9 - R_7 \\ -\frac{R_1}{0.14} - \frac{R_2}{0.5} \\ \frac{0.01}{0.14}R_1 - \phi\frac{R_3}{0.55} \\ 0 \end{bmatrix} ; \; g(x) = \begin{bmatrix} -\frac{E}{V} \\ -\frac{X_1}{V} \\ -\frac{X_2}{V} \\ -\frac{X_3}{V} \\ \frac{S_{in} - S}{V} \\ -\frac{A}{V} \\ 1 \end{bmatrix} ; \; h(x) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ S \\ 0 \\ 0 \end{bmatrix} \tag{19.36}$$

x, y, u, f and g are respectively the state, output, control input, vector field of the dynamics and vector field of the control.

Relative Degree of the System

The relative degree of a nonlinear system is defined by (17.55).

In the present case, the Lie derivatives of the output $y = S$ are

$$
\begin{aligned}
L_f^0 S &= S \\
L_f^1 S &= \frac{\partial S}{\partial x} f \\
&= \begin{bmatrix} \frac{\partial S}{\partial E} & \frac{\partial S}{\partial X_1} & \frac{\partial S}{\partial X_2} & \frac{\partial S}{\partial X_3} & \frac{\partial S}{\partial S} & \frac{\partial S}{\partial A} & \frac{\partial S}{\partial V} \end{bmatrix}
\begin{bmatrix} \frac{0.45}{0.14} R_1 - (1-\phi)\frac{R_3}{0.55} \\ R_1 + R_5 + R_9 - R_4 - R_8 \\ R_2 + R_4 + R_7 - R_5 - R_6 \\ R_3 + R_8 + R_6 - R_9 - R_7 \\ -\frac{R_1}{0.14} - \frac{R_2}{0.5} \\ \frac{0.01}{0.14} R_1 - \phi \frac{R_3}{0.55} \\ 0 \end{bmatrix} \\
&= -\frac{R_1}{0.14} - \frac{R_2}{0.5}
\end{aligned}
\tag{19.37}
$$

$$
\begin{aligned}
L_g L_f^0 S &= L_g S \\
&= \frac{\partial S}{\partial x} g \\
&= \begin{bmatrix} \frac{\partial S}{\partial E} & \frac{\partial S}{\partial X_1} & \frac{\partial S}{\partial X_2} & \frac{\partial S}{\partial X_3} & \frac{\partial S}{\partial S} & \frac{\partial S}{\partial A} & \frac{\partial S}{\partial V} \end{bmatrix}
\begin{bmatrix} -\frac{E}{V} \\ -\frac{X_1}{V} \\ -\frac{X_2}{V} \\ -\frac{X_3}{V} \\ \frac{S_{in} - S}{V} \\ -\frac{A}{V} \\ 1 \end{bmatrix} \\
&= \frac{S_{in} - S}{V} \neq 0 \quad \text{as} \quad S_{in} \neq S
\end{aligned}
\tag{19.38}
$$

The relative degree of system (19.36) is thus $r = 1$.

Synthesis of the Nonlinear Control Law

For a system of relative degree equal to 1, according to Isidori (1995), the state feedback (Eq. 17.101), which makes the closed-loop input-output behaviour $v - y$ linear, is

$$
u = \frac{\dot{z}_1 - L_f h(x)}{L_g h(x)} \tag{19.39}
$$

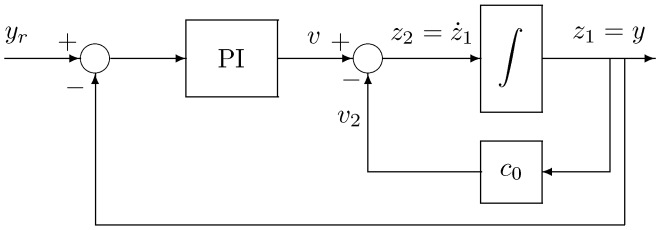


Figure 19.11: Nonlinear geometric control of the biological reactor with pole-placement

In fact, we realize a pole-placement (Fig. 19.11) by introducing a feedback (Eq. 17.119) such that

$$u = \frac{v - c_0 h(x) - L_f h(x)}{L_g h(x)} \tag{19.40}$$

where c_0 is a constant scalar parameter. Again, it is necessary to modify Eq. (19.40) so as to take into account the reference trajectory as in Eq. (17.133).

The external input v is defined by means of a PI controller, which allows us to eliminate the stationary error eventually due to the modelling errors and the unmeasured disturbances, etc.

$$v = K_c \left[(y_r(t) - y(t)) + \frac{1}{\tau_I} \int_0^t (y_r(\tau) - y(\tau)) d\tau \right] \tag{19.41}$$

where K_c and τ_I are respectively the proportional gain and the integral time constant. The transfer function of the closed-loop system is then the following

$$\frac{Y(s)}{Y_r(s)} = \frac{K_c s + K_c/\tau_I}{s^2 + (c_0 + K_c)s + K_c/\tau_I} \tag{19.42}$$

The scalar parameters K_c, τ_I, c_0 are chosen in order to approach a polynomial minimizing an ITAE criterion; we verify that the following polynomial is Hurwitz (the poles have a negative real part) to ensure the closed-loop stability related to the roots of the characteristic equation

$$s^2 + (c_0 + K_c)s + K_c/\tau_I = 0 \tag{19.43}$$

19.3.4 Simulation Conditions

The results have been obtained in simulation. The fermentation is realized in a fed-batch reactor, fed with glucose at concentration $S_{in} = 300$ g/l (Fig. 19.9). The initial conditions are

$$\begin{array}{llll} E = 0.0 \text{ g/l}\,, & X_1 = 3.0 \text{ g/l}\,, & X_2 = 0.0 \text{ g/l}\,, & X_3 = 0.0 \text{ g/l} \\ S = 0.0 \text{ g/l}\,, & A = 0.0 \text{ g/l}\,, & V = 10 \text{ l} & \end{array} \tag{19.44}$$

An actual fermentation is performed in two stages. The first stage is run in batch mode and the yeast first consumes the glucose initially present, then

the ethanol that it produces. When the ethanol concentration becomes close to zero, the closed-loop control can start, the glucose feed is realized and the fermentation is then performed in fed-batch mode.

The maximum volume maximal of the reactor is equal to 20 l. The maximum acceptable feed flow rate is 5 l/h.

In the case of yeast production, the set point is a glucose concentration equal to $S_c = 0.07$ g/l. Indeed previous studies have shown that this set point value is nearly optimal for maximizing the biomass productivity (Dantigny, 1989).

In order to simulate a real experience, to the ethanol and glucose measurements, a Gaussian noise of respective deviations 5×10^{-3} g/l for ethanol and 5×10^{-3} g/l for glucose is added to the value provided by the theoretical model.

At the initial instant, the glucose concentration is zero; immediately afterwards, the glucose concentration set point must follow a step equal to 0.07 g/l. To avoid rapid variations of the control input, the set point was filtered by an overdamped second-order filter ($\tau = 5$, $\zeta = 1.2$), so that in reality the output must follow this reference trajectory.

In the case of a real fermentation, the sampling period of the glucose concentration measurement is 15 minutes. The actual value of this glucose concentration in the reactor is obtained about 15 minutes after the corresponding sampling. However, a measurement of the ethanol concentration is possible every 3 minutes by means of a volatile component sensor. With the control variable, i.e. the glucose feed flow rate, being calculated every 3 seconds, a state observer is necessary to reconstruct the measured states as well as the unmeasured ones (Bastin, 1988; Bastin and Dochain, 1990; Bastin and Joseph, 1988; Lucena et al., 1995; Watanabe, 1992) based on these measurements. In the present case, the extended Kalman filter is used with the nonlinear process model. Some states are initially supposed to be known with a given error. The initialization of the estimations was taken as

$$\begin{array}{llll} \hat{E}_0 = 1.0 \text{ g/l} , & \hat{X}_{10} = 3.1 \text{ g/l} , & \hat{X}_{20} = 0.0 \text{ g/l} , & \hat{X}_{30} = 0.0 \text{ g/l} \\ \hat{S}_0 = 1.10^{-3} \text{ g/l} , & \hat{A}_0 = 0.0 \text{ g/l} , & \hat{V}_0 = 10 \text{ l} & \end{array} \tag{19.45}$$

The volume is perfectly known and for this reason is not estimated.

The parameters used for the control law are

$$c_0 = 5.655 \qquad K_c = 0.628 \qquad \tau_I = 0.0637\text{h} \tag{19.46}$$

To avoid the eventual windup effects due to the integral action present in the control law, when the deviation between set point and measurement, taken as an absolute value, is larger than a given value (here 0.04 g/l), the integral action is cut. It is reintroduced when the deviation becomes smaller than the fixed threshold.

In a real process, we avoid modifying the position of the valve too frequently. A period of constant value equal to 3.6 s was assumed for the fermentor.

19.3.5 Simulation Results

The glucose concentration (Fig. 19.12) starts from 0 g/l and increases up to its set point with little overshoot, following its reference trajectory. In the following, the value of the measured output oscillates between 0.06 and 0.08 g/l. It is possible to verify that the estimation always remains in the confidence interval of the measured value. The oscillations visible at the end are probably due to the bad estimation of acetate, whose influence in the model begins to become sensitive.

The ethanol concentration (Fig. 19.13) starts from 0 g/l and goes through a maximum of around 0.4 g/l before going back to 0.07 g/l, corresponding to the stoichiometric transformation of glucose into ethanol. This is favourable for good yeast productivity.

The total biomass concentration ($X_1 + X_2 + X_3$) is represented in Fig. 19.14. After 15 hours of fermentation, the biomass concentration reaches 33 g/l. The desired productivity could even be more important. In the present case, a time limitation was imposed on the fermentation. If the latter were to continue, the exponential allure of biomass growth would be maintained, but a volume limitation would then be imposed. Beyond some biomass concentration, practical problems such as oxygenation of the fermentation medium occur.

The evolution of the manipulated glucose feed flow rate is presented in Fig. 19.15. The flow rate increases in a slow and regular manner. A remarkable characteristic is the exponential allure of the control profile which demonstrates the highly nonlinear behaviour of this system and renders classical linear control very difficult. The flow rate remains largely lower than its maximum authorized value (30 l/h). The evolution of the liquid volume (Fig. 19.16) of course follows that of the feed flow rate. The constraint of maximum volume is not reached after 15 h.

19.3.6 Conclusion

The nonlinear control law designed in the framework of differential geometry allows us to fully exploit the knowledge of a complex process, presently the production of *Saccharomyces cerevisiae* yeast during a fed-batch fermentation. The desired objective is to maximize the productivity.

The obtained control presents very good properties, although the state-space model of the process is highly nonlinear and non-stationary.

One important problem is the synthesis of the state observer necessary for reconstructing the unmeasured states. The met difficulties are, in particular, important modelling uncertainties, inaccurate measurements, different sampling periods and measurement delays (Lucena et al., 1995).

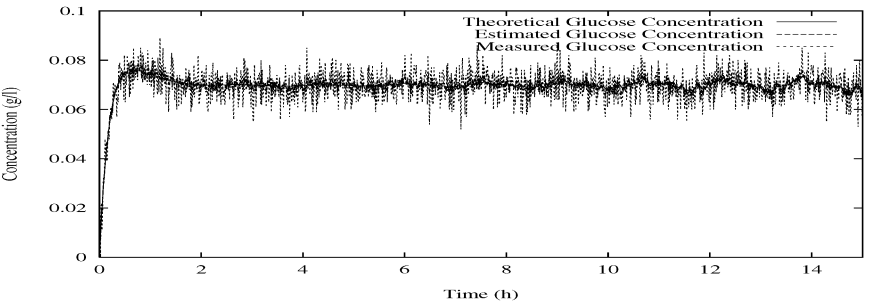


Figure 19.12: Measured and estimated glucose concentration

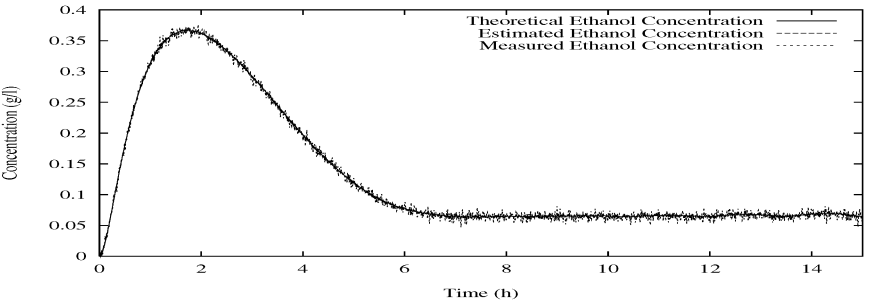


Figure 19.13: Measured and estimated ethanol concentration

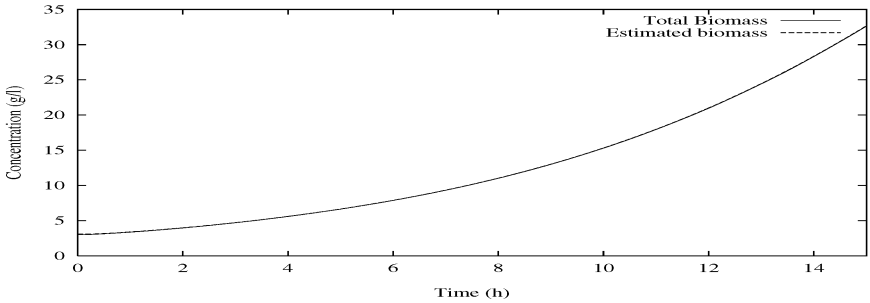


Figure 19.14: Theoretical and estimated biomass concentration

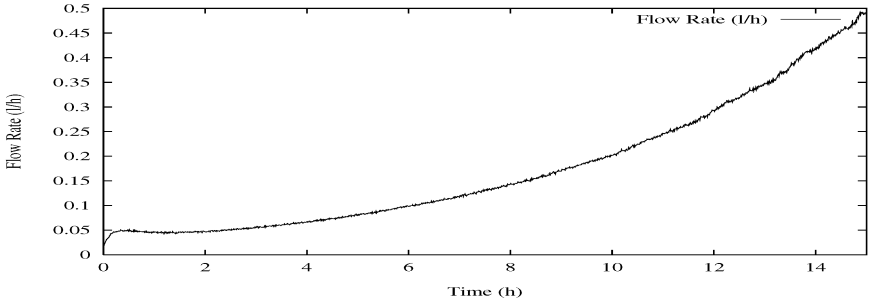


Figure 19.15: Manipulated variable: glucose flow rate.

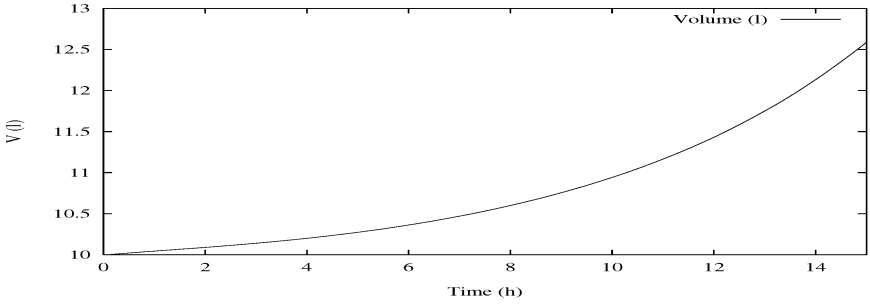


Figure 19.16: Volume of the biological reactor

References

G. Bastin. State estimation and adaptative control of multi-linear comportmental system : Theoretical framework and application to (bio)chemical processes. In *CNRS Symposium on Nonlinear Systems*, Nantes, France, 1988.

G. Bastin and D. Dochain. *On-Line Estimation and Adaptive Control of Bioreactors*. Elsevier, Amsterdam, 1990.

G. Bastin and P.D. Joseph. Stable adaptative observers for nonlinear time varying systems. *IEEE Trans. Automat. Control*, 33:650–658, 1988.

P. Dantigny. *Cinétique, Modélisation de la Croissance de* Saccharomyces cerevisiae*, Commande Non Linéaire de Type L/A d'un Procédé Semi-Continu*. Phd thesis, INPL, Nancy, France, 1989.

P. Daoutidis, C. Kravaris, and M. Soroush. Feedforward/feedback control of multivariable nonlinear processes. *AIChE J.*, 36(10):1471–1484, 1990.

C. Gentric, F. Pla, M.A. Latifi, and J.P. Corriou. Optimization and non-linear control of a batch emulsion polymerization reactor. *Chem. Eng. J.*, 75:31–46, 1999.

K.A. Hoo and J.C. Kantor. Linear feedback equivalence and control of an unstable biological reactor. *Chem. Eng. Commun.*, 46:385–399, 1986.

A. Isidori. *Nonlinear Control Systems: An Introduction*. Springer-Verlag, New York, 2nd edition, 1989.

A. Isidori. *Nonlinear Control Systems*. Springer-Verlag, New York, 3rd edition, 1995.

C. Kravaris, R.A. Wright, and J.F. Carrier. Nonlinear controllers for trajectory tracking in batch processes. *Comp. Chem. Engng.*, 13(1/2):73–82, 1989.

L.C. Limqueco and J.C. Kantor. Nonlinear output feedback control of an exothermic reactor. *Comp. Chem. Eng.*, 14:427–437, 1990.

S. Lucena, C. Fonteix, I. Marc, and J.P. Corriou. Nonlinear control of a discontinuous bioreactor with measurement delays and state estimation. In A. Isidori, editor, *European Control Conference ECC95*, volume 4, pages 3811–3815, Rome, Italie, 1995.

S. Lucena, A.M. Souto Maior, C. Fonteix, I. Marc, and J.P. Corriou. Application of nonlinear control on a fermentation process: production of *saccharomyces cerevisiae*. *Latin American Applied Research*, 31:235–240, 2001.

W. L. Luyben. *Process Modeling, Simulation, and Control for Chemical Engineers*. McGraw-Hill, New York, 1990.

T. Pröll and N.M. Karim. Nonlinear control of a bioreactor model using exact and I/O linearization. *Int. J. Control*, 60(4):499–519, 1994.

A. Rajab. *Un Modèle Physiologique de la Croissance de* Saccharomyces cerevisiae *en Fermenteur Continu et Discontinu*. Phd thesis, INPL, Nancy, France, 1983.

M. Soroush and C. Kravaris. Nonlinear control of a batch polymerization reactor: an experimental study. *AIChE J.*, 38(9):1429–1448, 1992.

M. Soroush and C. Kravaris. Multivariable nonlinear control of a continuous polymerization reactor: an experimental study. *AIChE J.*, 39(12):1920–1937, 1993.

M. Soroush and C. Kravaris. Nonlinear control of a polymerization CSTR with singular characteristic matrix. *AIChE J.*, 40(6):980–990, 1994.

K. Watanabe. *Adaptive Estimation and Control*. Prentice Hall, London, 1992.

Chapter 20

Distillation Column Control

Distillation columns are unit operations that are very common in chemical, petrochemical and even sometimes in metallurgical industries. Moreover, they consume a large part of a plant's total energy. The optimization of their design and operation is thus an essential objective. The more and more severe operation constraints which are imposed make their mastering more delicate and implies impressive control strategies.

20.1 Generalities for Distillation Columns Behaviour

The purpose of a distillation column is to separate a multicomponent feed into products of different compositions or to purify intermediate or final products. The possibility of distillation relies on the volatility difference existing between different chemical species. If we consider a simple vessel containing a mixture of two components (binary mixture) and if its contents are heated, there occurs an equilibrium ruled by thermodynamics between liquid and vapour which then have different compositions: the vapour is richer in the more volatile component ("light") and the liquid richer in the less volatile component ("heavy"). In a first approximation, the most volatile component is that which possesses the lowest molecular weight. Note that in the case of a multicomponent mixture, such as a mixture of hydrocarbons, the prediction of the respective volatilities is more delicate, but can be performed by most thermodynamic actual codes (Prausnitz and Chueh, 1968; Prausnitz et al., 1967). Through this equilibrium, a separation operation called flash is realized, which can be considered as the operation that occurs at the level of each tray of a distillation column. When the separation by a simple flash is insufficient, the separation is operated in a distillation column which will allow better separation due to the series of trays. A packed column, in fact, behaves in an analogous manner to a real tray

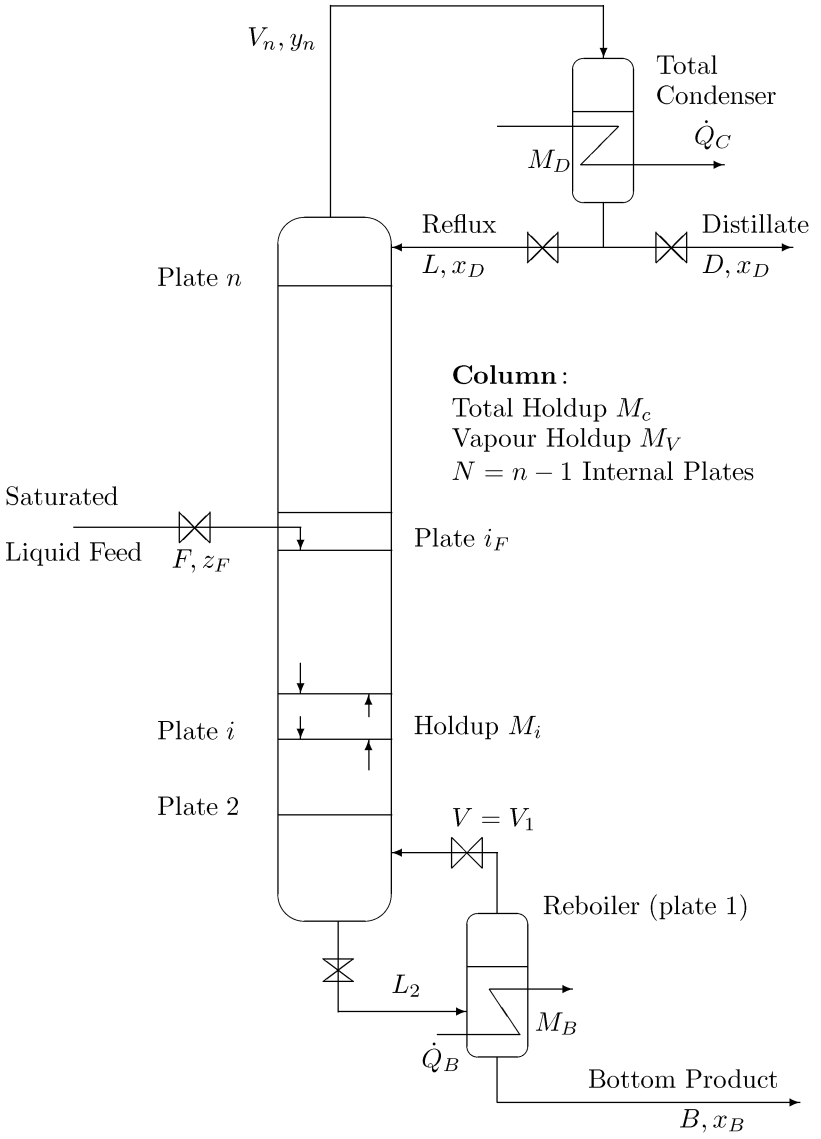


Figure 20.1: Distillation column

column, but in that case the notion of theoretical separation stage is used.

The technology of distillation columns can be complex and this review is limited to the classical case of a column having only one feed, producing the bottom product at the reboiler and the distillate or overhead product at the condenser (Fig. 20.1), without any side withdrawal. The distillate can be seen as the production of "light" and the bottom product as the production of

"heavy". A multicomponent mixture will thus be considered as being composed of two key-components, a light component and a heavy component, as if it were a binary system.

Many possibilities for more complex distillation are used. Thus, the distillation columns can have several side streams, allowing us to realize intermediate cuts (petroleum distillation). There can exist extractive distillation columns (additional solvent feed) (Lang et al., 1995) or reactive distillation (with chemical reaction) (Albet et al., 1991). On the other hand, there exist batch distillation columns in which the feed is introduced into the reboiler at the initial instant, and for which a control policy must be defined to allow for optimal separation of the products with desired purities at different instants.

Here, we are concerned with a distillation column that we can qualify as classical and representative of a large number of industrial columns (Fig. 20.1). The role of the reboiler situated at the bottom of the column is to bring to the whole column the energy $\dot{Q}_B$ necessary for the separation operation, corresponding to the vapourization enthalpy. The condenser can be total or partial, depending on whether it condenses the totality or a part of the vapour arriving at the top of the column into a liquid that is separated between the distillate and the reflux. In the simplest case constituted by the total condenser, the absorbed heat $\dot{Q}_C$ (supply of cold) allows us to exactly perform the condensation of the head vapour. The reboiler and the condenser are, in fact, heat exchangers. The feed F can be introduced into the column at different enthalpy levels (subcooled liquid, saturated liquid, liquid-vapour mixture, saturated vapour, overheated vapour). Frequently, the feed is a saturated liquid (at the boiling point), which will be our hypothesis.

A column is traditionally divided into two large sections: above the feed tray, the rectifying section and below the feed tray, the stripping section. In general, a column is operated at a fixed top pressure in order to ensure an effective liquid-vapour equilibrium on all the plates. A temperature gradient is then established in the column, the reboiler being at the highest temperature and the top plate at the lowest temperature.

A column tray can be realized in different ways, e.g. a valve tray, bubble cap tray, sieve tray (Rose, 1985). The tray is provided with a weir (Fig. 20.2). A given liquid level coming from the downcomer is thus maintained on the plate. The vapour coming from the lower plate pushes on the valves (for a valve tray) and crosses the plate liquid, thus enriching itself in volatile components. On the other hand, the liquid passing over the weir is enriched in heavy components and flows down to the lower plate.

The column is thus crossed by a downward liquid stream and an upward vapour stream. The plates of the column are numbered in ascending order: the reboiler is 1, and the top plate n. The condenser can only be considered as a theoretical stage if it is partial, i.e. if it realizes a liquid-vapour equilibrium and for this reason is considered separately. Any plate i is thus modelled (Fig. 20.3) by means of entering and exiting streams; those entering: the liquid L_{i+1} comes from the upper plate $i+1$ and the vapour V_{i-1} comes from the lower plate $i-1$; those exiting: the liquid L_i and the vapour V_i have the number

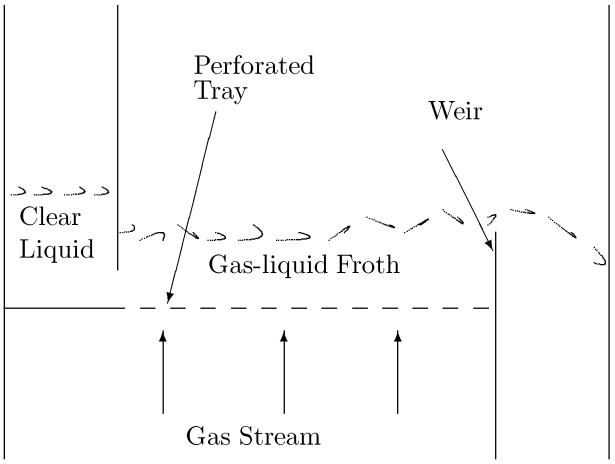


Figure 20.2: Distillation column plate

of the plate and are considered to be in equilibrium in the modelling. In the calculation of a real column, an efficiency coefficient (Murphree efficiency) is introduced to take into account the deviation with respect to equilibrium.

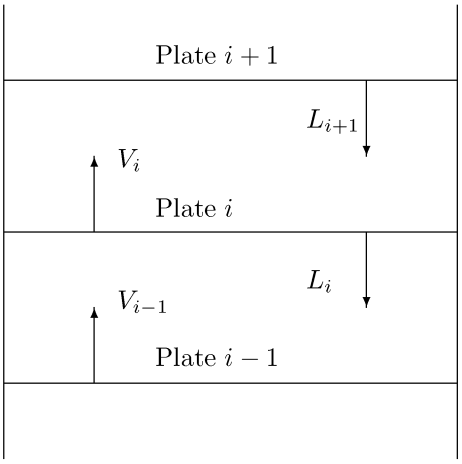


Figure 20.3: Plate model

20.2 Dynamic Model of the Distillation Column

The modelling of a distillation column can be performed at different complexity levels with regard to the objective. Models called rigorous (Skogestad, 1992, 1997) concern models which include the mass and energy balances at each

stage, a dynamic model of the liquid flow and a model of the pressure dynamics, possibly a model of the reboiler and the condenser. In fact, even in a rigorous model, simplifications are introduced, such as the perfect mixing at each stage and the thermal and thermodynamic equilibrium between phases (a Murphree efficiency per plate may be introduced to correct for non-equilibrium).

With the considered distillation having n_c components, the equations for each theoretical stage i of the column include:

- The mass balance per component j (n_c equations):

$$\frac{dM_i\, x_{i,j}}{dt} = L_{i+1}\, x_{i+1,j} + V_{i-1}\, y_{i-1,j} - L_i\, x_{i,j} - V_i\, y_{i,j} \tag{20.1}$$

- The energy balance (1 equation):

$$\frac{dU_i}{dt} = h^l_{i+1}\, L_{i+1} + h^v_{i-1}\, V_{i-1} - h^l_i\, L_i - h^v_i\, V_i \tag{20.2}$$

- The thermodynamic equilibrium for each component j (n_c equations):

$$\mu^l_j(T_i, P_i, x_{i,j}) = \mu^v_j(T_i, P_i, y_{i,j}) \Longleftrightarrow y_{i,j} = k(x_{i,j}) \tag{20.3}$$

where $x_{i,j}$ and $y_{i,j}$ respectively indicate the liquid and vapour mole fractions in component j at plate i. L_i and V_i are respectively the liquid and vapour molar flow rates leaving plate i. h^l and h^v are respectively the liquid and vapour enthalpies. μ^l_j and μ^v_j are respectively the liquid- and vapour-phase chemical potentials in component j. The global mass balance results from the n_c mass balance equations per component.

The number of accumulated moles M_i and the accumulated internal energy U_i are calculated by using the description of the liquid holdup on the plate, thus its hydrodynamic behaviour. Gani et al. (1986) give many hydrodynamic correlations that allow us to calculate the liquid and vapour flow rates, the pressure drop, the conditions of flooding, weeping, of liquid and vapour entrainment, for different types of plates.

Besides any plate i, the feed plate, the reboiler and the condenser (top drum) must be described according to analogous equations, but with taking particularities into account. For a detailed description, it is possible to refer to many general books of chemical engineering (Holland, 1981). However, few books treat dynamic problems in a detailed manner.

A rigorous description allows us to obtain a dynamic reference model. In many physical cases, important simplifications are brought to the previous model. Skogestad (1997) details the implications of the modelling simplifications, concerning in particular the vapour dynamics, energy balance, and liquid flow dynamics. The simplified model can also be used by the control specialist as a nonlinear model in view of the design of a control law.

The model described below is not a rigorous model and the differences with a rigorous model are underlined. On the contrary, it is well adapted to nonlinear control (Creff, 1992). We consider the column as a binary column, separating a heavy component and a light component. The retained hypotheses are:
a) Of thermodynamic order:

– The liquid and the vapour are homogeneous and at thermodynamic equilibrium on each plate.
– The partial molar enthalpy of a component j is independent of pressure, temperature and composition, thus is constant in each phase for any plate i. This can be applied when the components have close chemical properties.
– The partial molar enthalpies of vapourization are identical for all components.

b) concerning the column:

– The pressure is constant on each plate, thus the hydrodynamic model becomes useless.
– The molar liquid holdup is constant on each plate.

From all these hypotheses, it results that the molar flow rates L_i and V_i will be constant in each section of the column, thus on each side of the feed, in steady-state regime as well as in dynamic regime. In the following, they are respectively denoted by L and V.

An important consequence is that the energy balance equation becomes useless. There remain the mass balance equations, which take simple expressions as the feed is considered to be binary. The expressions are written with respect to the most volatile component. The simplified model is written keeping the hypotheses of the total condenser and the saturated liquid feed:
Condenser ($i = n + 1 = D$):

$$M_D \frac{dx_D}{dt} = V\, y_n - V\, x_D \tag{20.4}$$

Rectification plates ($i = i_F + 1, \ldots, n$):

$$M_i \frac{dx_i}{dt} = L\, x_{i+1} + V\, y_{i-1} - L\, x_i - V\, y_i \tag{20.5}$$

Feed plate ($i = i_F$):

$$M_{i_F} \frac{dx_{i_F}}{dt} = L\, x_{i_F+1} + V\, y_{i_F-1} + F\, z_F - (L + F)\, x_{i_F} - V\, y_{i_F} \tag{20.6}$$

Stripping plates ($i = 2, \ldots, i_F - 1$):

$$M_i \frac{dx_i}{dt} = (L + F)\, x_{i+1} + V\, y_{i-1} - (L + F)\, x_i - V\, y_i \tag{20.7}$$

Reboiler ($i = 1 = B$):

$$M_1 \frac{dx_1}{dt} = (L + F)\, x_2 - (L + F - V)\, x_1 - V\, y_1 \tag{20.8}$$

It must be noted that in these equations, on each plate, the vapour phase mole fraction y_i is a nonlinear function of the liquid phase mole fraction x_i, calculable by the relations of thermodynamic equilibrium (Prausnitz et al., 1967).

Many papers deal with the numerical aspects related to the solving of distillation models. Indeed, a rigorous dynamic distillation model includes a large set of ordinary differential and algebraic equations. The problem formulation has important consequences on its solving (Cameron et al., 1986). The problem structure makes a hollow and band matrix appear. Moreover, Cameron et al. (1986) distinguish two modes of simulation according to whether the column is operated in continuous mode (in the neighbourhood of a stationary point) or in discontinuous mode (startup or shutdown or feed change (Ruiz et al., 1988)). In discontinuous mode, the transients are much more severe and sequential operations may occur. In continuous mode, however, we must be cautious about the flooding and weeping hazards inducing operating discontinuities. Cameron et al. (1986) evaluate the necessity of using numerical integration schemes of stiff systems. Generally, semi-implicit schemes of the Runge-Kutta type or totally implicit, based in general on Gear's method, are recommended (Gallun and Holland, 1982; Gear and Petzold, 1984; Hindmarsh and Petzold, 1995; Unger et al., 1995).

Although the model of a packed column, either for distillation or absorption, is a system of partial differential equations, it is common to model such a column by using the notion of theoretical stage, and the problem becomes identical to that of a distillation column with theoretical plates. It must be noted that the dynamics of a packed distillation column is neatly faster than that of a plate column, as the liquid holdup is lower (frequently by a factor of two or three) (Skogestad, 1992).

Skogestad and Morari (1988) show that the dynamic behaviour of a distillation column is that of a two time-constant system. The composition response to external flow rate variations (which makes the ratio D/B vary) is close to that of a first-order system with time constant τ_1, which can be very large if the exit products are very pure. Consider the influence of external flow rate variations (F or D or B) on the global mass balance of the column. Assuming that the molar holdup is constant for any plate and that the column goes from one initial steady state to a final steady state (subject to a variation Δ), the time constant τ_c is equal to

$$\tau_c = \frac{\sum_{i=1}^{n+1} M_i \Delta x_i}{D_f \Delta y_D + B_f \Delta x_B} \tag{20.9}$$

This time constant τ_c is the contribution of the holdup inside the column, in the reboiler and the condenser. Equation (20.9) does not correspond to a linearization and allows us to obtain an excellent agreement with the nonlinear responses, even for high-purity columns. For binary columns, the linearized formula

$$\tau_{c,l} = \frac{M_c}{[Bx_B(1-x_B) + Dy_D(1-y_D)] \ln S} \qquad \text{with: } S = \frac{y_D(1-x_B)}{(1-y_D)x_B} \tag{20.10}$$

provides a good approximation of τ_c while demonstrating the purity influence through the separation factor S. The response to internal flow rates variations

(which modify L and V while keeping D and B constant) is also first-order, but its time constant τ_2 is much smaller than τ_1; however, its influence is important. The time constant τ_2 corresponds to a variation of internal flows and is near M_c/F. Thus, it is possible to correctly represent the dynamic behaviour of composition responses by an analytical model using two time constants. Moreover, the use of the top logarithmic compositions, $\log(1 - y_D)$, and bottom ones, $\log x_B$, nearly totally cancels the nonlinearity. The introduction of feedback to control the column modifies the poles and the closed-loop responses can be largely faster than could be thought from the time constants τ_1 and τ_2 corresponding to open-loop responses.

A reduced linear model is often sought for control; however, the used simplifications can be excessive. Skogestad (1992) recommends realizing a model reduction from a linearized model which is itself deduced from a complete rigorous model.

20.3 Generalities for Distillation Column Control

The most common objective in a distillation process is to maintain the top and bottom compositions at a desired specification.

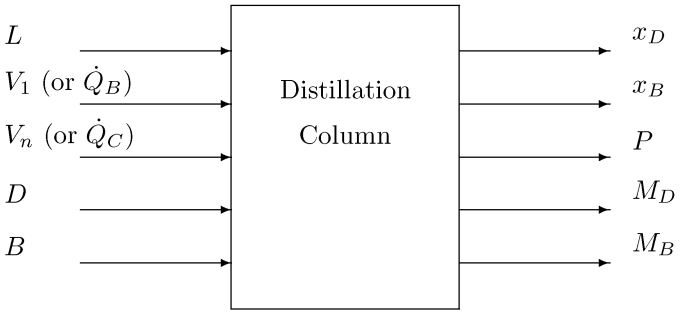


Figure 20.4: Block diagram of a distillation column

A typical distillation column (Fig. 20.1) can be represented as a block diagram (Fig. 20.4) which possesses five control inputs u corresponding to five valves (flow rates of distillate D, bottom product B, reflux L, reboiler vapour V_1 (indirectly manipulated by the heat power at the reboiler $\dot{Q}_B$), top condensed vapour V_n (indirectly manipulated by the power withdrawn from the condenser) and five controlled outputs y. Three of the controlled variables (holdup at the reboiler M_B and at the condenser M_D, and pressure P or vapour holdup M_V) must be carefully controlled to maintain the stability of the operation. In fact, Jacobsen and Skogestad (1994) show that even binary columns whose pressure and levels are controlled can present multiple stationary states. There remain two degrees of freedom for the top and bottom compositions x_D and x_B. The

other inputs are those which are not affected inside the system: the disturbances d and the set points y_r. The disturbances in the column, in general, are related to the feed (flow rate F, feed enthalpy expressed with respect to the liquid fraction k_F and the feed composition z_F). In a general manner, disturbances are classified as measured disturbances (e.g. often F) and unmeasured (e.g. often z_F). The set points can change, for example, when an on-line economic optimization is performed. The outputs can be measured (pressure P, holdup M_B and M_D, top and bottom compositions with time delay); besides, the temperature is generally measured at several locations. For the distillation column, we can consider that the state variables are the liquid and vapour mole fractions at the level of each plate; unfortunately, in general, they are not measured inside the column. The temperature measurement on sensitive plates allows us, by means of the thermodynamic equilibria and knowing the pressure, to estimate the profile of mole fractions.

The dynamics of the five outputs are of different orders, which allows hierarchical control. Indeed, the time constants of the pressure dynamics and the reboiler and condenser levels dynamics are smaller, i.e. around a few minutes. On the contrary, the time constants of the top and bottom mole fractions dynamics are largely higher, i.e. able to reach several hours.

Problems encountered by this coupled control of top and bottom mole fractions, increased in the case of high-purity columns, are:

- The strongly nonlinear behaviour of the process.
- An often slow response.
- Measurement problems (time delays for composition measurements).
- The difficult choice of the control variables for composition control.
- A highly interactive system.
- The system is badly conditioned, particularly for high-purity columns.

The example of the design of a control system for a distillation column is remarkable in order to show the importance and the respective domains of system analysis, modelling and identification, the choice of a control algorithm. This problem has been studied very often and many control approaches applied to a distillation column have been proposed along the past years. We will recall some of them here which seem particularly interesting with regard to the methodology and evolution that they represent. The range of proposed solutions may leave more than one engineer thoughtful. Successively, single-input single-output control, multivariable control by decoupling (1970), by transfer function matrix (1975), auto-adaptive control (1981), bilinear models (1978), model predictive control (since 1978) and nonlinear control (1990) will be reviewed.

20.4 Different Types of Distillation Column Control

20.4.1 Single-Input Single-Output Control

The simplest used approach consists of controlling only one composition, generally top y_D. This is a single-input single-output control and, in this case, a flow rate is manually controlled by the operator. Taking into account strict specifications, dual control, which is aiming to control both the top x_D and bottom x_B mole fractions, is desirable and will be presented in the following. According to Skogestad and Morari (1987a), dual control allows us to spare 10 to 30% energy by avoiding over-purification and out of norms product loss.

20.4.2 Dual Decoupling Control

One of the main difficulties in multivariable feedback control lies in the interaction of the control loops, e.g. a modification of the reboiler vapour flow rate V_1 to control the composition x_B also influences x_D (Fig. 20.5). Different solutions were proposed, e.g. using a ratio between the reflux and the condenser vapour to reduce interaction. Luyben (1970) studied the decoupling from a linear model of the column.

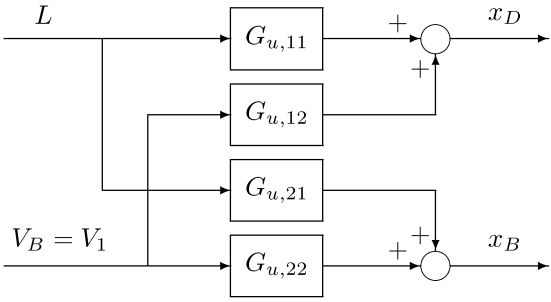


Figure 20.5: Interaction inside a distillation column

For distillation, we get a linear model expressing the outputs $\boldsymbol{Y}$ with respect to the inputs $\boldsymbol{U}$ and the disturbances $\boldsymbol{D}$

$$\boldsymbol{Y} = \boldsymbol{G}_u \boldsymbol{U} + \boldsymbol{G}_d \boldsymbol{D} \tag{20.11}$$

i.e.

$$\begin{bmatrix} x_D \\ x_B \end{bmatrix} = \begin{bmatrix} G_{u,11} & G_{u,12} \\ G_{u,21} & G_{u,22} \end{bmatrix} \begin{bmatrix} L \\ V_1 \end{bmatrix} + \begin{bmatrix} G_{d,11} & G_{d,12} \\ G_{d,21} & G_{d,22} \end{bmatrix} \begin{bmatrix} z_F \\ F \end{bmatrix} \tag{20.12}$$

We notice that x_D is simultaneously influenced by L and V_1, similarly for x_B with regard to L and V_1. The block diagram (Fig. 20.6) provides the following equation

$$\boldsymbol{Y} = [\boldsymbol{I} + \boldsymbol{G}_u \boldsymbol{G}_r]^{-1} \boldsymbol{G}_u \boldsymbol{G}_r \boldsymbol{Y}_r + [\boldsymbol{I} + \boldsymbol{G}_u \boldsymbol{G}_r]^{-1} \boldsymbol{G}_d \boldsymbol{D} \tag{20.13}$$

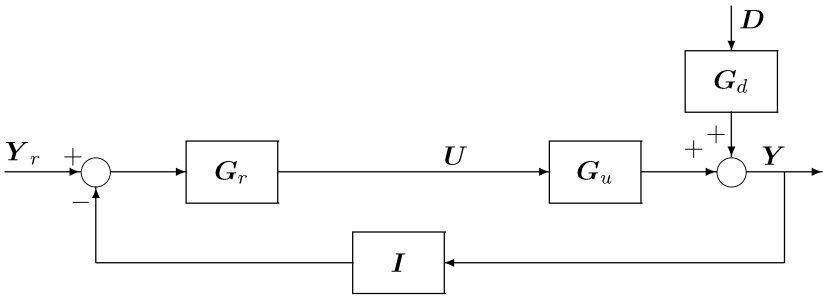


Figure 20.6: Closed loop without decoupling

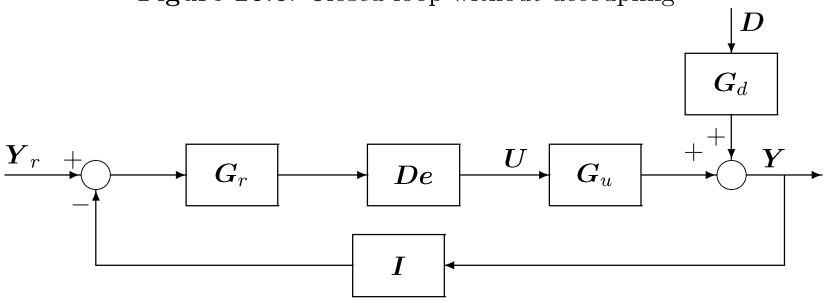


Figure 20.7: Closed loop with decoupling

To eliminate the interaction between the control loops, decoupling elements $\boldsymbol{De}$ are added which give the new scheme (Fig. 20.7) expressed by the equation

$$\boldsymbol{Y} = [\boldsymbol{I} + \boldsymbol{G}_u\,\boldsymbol{De}\,\boldsymbol{G}_r]^{-1}\boldsymbol{G}_u\,\boldsymbol{De}\,\boldsymbol{G}_r\,\boldsymbol{Y}_r + [\boldsymbol{I} + \boldsymbol{G}_u\,\boldsymbol{De}\,\boldsymbol{G}_r]^{-1}\,\boldsymbol{G}_d\,\boldsymbol{D} \qquad (20.14)$$

i.e.

$$\boldsymbol{Y} = \boldsymbol{K}_1\,\boldsymbol{Y}_r + \boldsymbol{K}_2\,\boldsymbol{D} \qquad (20.15)$$

For the control to be non-interactive, the matrix $\boldsymbol{K}_1$ of the closed-loop transfer function must be diagonal and specified. Luyben deduces the decoupling matrix

$$\boldsymbol{De} = [\boldsymbol{G}_u]^{-1}\,[\operatorname{diag}\boldsymbol{G}_u] \qquad (20.16)$$

According to Luyben, the realized experiences show the efficiency of this decoupling. An application to the linear model of distillation column (Wood and Berry, 1973) is presented in Example 8.1.

However, a noticeable drawback of the decoupler thus calculated is that it does not take into account the disturbances, e.g. in front of a feed flow rate disturbance, the decoupled system behaves in a worse manner compared to the undecoupled system.

On the contrary, Niederlinski (1971) shows that it can be interesting to use interaction because it can offer a better disturbance alleviation than the non-interactive design. Moreover, in some cases, Luyben's decoupler can be unrealizable (Weischedel and McAvoy, 1980).

A robustness analysis by singular value decomposition concerning decoupling control of distillation columns (Sect. 8.6) was realized (Arkun and Morgan, 1988; Arkun et al., 1984). For high-purity columns, the system is badly conditioned, i.e. the singular values have very different orders of magnitude. This property explains many characteristics, and Skogestad (1992) strongly advises against using decoupling when the model has RGA matrix elements that are large compared to 1. Control has been analyzed in the robustness context (Skogestad and Lundstrom, 1990; Skogestad et al., 1988).

20.4.3 The Column as a 5 × 5 System

Skogestad and Morari (1987a) show that although a distillation column is strongly nonlinear, for control needs it can be described by a linear model. The composition responses are much slower than those of the top and bottom flow rate, which makes the open-loop transfer functions difficult to obtain by simulation. However, as the composition response depends very little on the levels in the reboiler and the condenser, these levels are often considered constant to find the response. Pressure has an important effect on composition, but is well regulated. With these approximations, the structure of the open-loop transfer matrix is given in Table 20.1.

Table 20.1: The distillation column seen as a 5 × 5 system

Controlled output	Manipulated variable				
	L	$V = V_1$	D	B	V_n
x_D	$G_{u,11}(s)$	$G_{u,12}(s)$	0	0	0
x_B	$G_{u,21}(s)$	$G_{u,22}(s)$	0	0	0
M_D	$-\frac{1}{s}$	0	$-\frac{1}{s}$	0	$\frac{1}{s}$
M_B	$\frac{1}{s}\exp(-\theta s)$	$\frac{-1-\lambda(1-\exp(-\theta s))}{s}$	0	$-\frac{1}{s}$	0
M_V	0	$\frac{1}{s+k_p}$	0	0	$-\frac{1}{s+k_p}$

The following hypotheses were retained:

- Constant molar flows.
- The transfer function for M_V is not a pure integrator because of the condensation effects included in k_p.
- $\exp(-\theta s)$ with $\theta = \tau_L\, N$ is an approximation for $1/(1+\tau_L s)^N$, $\tau_L = (\partial M_i/\partial L_i)_{V_i}$ is the hydraulic time constant, N being the total number of plates.
- $\lambda = (\partial L_i/\partial V_i)_{M_i}$ is the variation of liquid flow rate L_i due to a variation of vapour flow rate V_i.
- Moreover, dynamics will occur in particular for the valves modifying L, V, D, B and V_n.

We notice from the matrix of Table 20.1 that the outputs x_D and x_B depend only on the inputs L and V through a 2×2 sub-matrix $\boldsymbol{G}$. Similar transfer matrices will occur for the disturbances.

In general, a 5×5 controller will not be used, but a 2×2 controller denoted by K for composition control, plus a control system for levels and pressure. If the flow rates L and V are chosen as control variables for composition control, the control configuration LV is represented by the following matrix

$$\begin{bmatrix} dL \\ dV \\ dD \\ dB \\ dV_n \end{bmatrix} = \begin{bmatrix} K & \dots & 0 & 0 & 0 \\ \dots & \dots & 0 & 0 & 0 \\ 0 & 0 & c_D(s) & 0 & 0 \\ 0 & 0 & 0 & c_B(s) & 0 \\ 0 & 0 & 0 & 0 & c_V(s) \end{bmatrix} \begin{bmatrix} dx_D \\ dx_B \\ dM_D \\ dM_B \\ dM_V \end{bmatrix} \tag{20.17}$$

The flow rate inputs L and V depend simultaneously on the compositions x_D and x_B, whereas the inputs D, B, V_n depend on only one output, respectively the holdups M_D, M_B and M_V.

Instead of simply using L and V as control variables, the ratio L/V (which is the slope of the operating line in MacCabe and Thiele diagram) and V/B can be used. This recommended configuration (Shinskey, 1984) corresponds, in fact, to a nonlinear control scheme. The controller becomes

$$\begin{bmatrix} dL \\ dV \\ dD \\ dB \end{bmatrix} = \begin{bmatrix} K & \dots & (R/D)_{c_D} & 0 \\ \dots & \dots & 0 & (V/B)_{c_B} \\ 0 & 0 & c_D & 0 \\ 0 & 0 & 0 & c_B \end{bmatrix} \begin{bmatrix} dx_D \\ dx_B \\ dM_D \\ dM_B \end{bmatrix} \tag{20.18}$$

We notice that the flow rates L and V depend both on compositions (x_D and x_B) and on levels (M_D for L and M_B for V). When the holdup M_D varies, the controller modifies both L and D, similarly for M_B and V and B. Each of the SISO level controllers modifies two flow rates.

Many control configurations are possible; among the factors that guide the choice (Skogestad and Morari, 1987a) are:

- The uncertainty: the analysis by relative gain array (RGA) (Shinskey, 1984; Skogestad and Morari, 1987b), allows to display the input-output couplings. Originally, RGA is a measure of steady-state interaction and depends on the process model, thus is independent of the control. It has been extended to dynamic interaction. A drawback of RGA analysis is that it does not take into account the disturbances. Analogous methods considering the disturbances can be used in closed loop (Hovd and Skogestad, 1992).
- The importance attached to set point tracking with respect to the influence of disturbances.
- The dynamic considerations (level loops influencing the flow rates L and V).
- The flow rate disturbance rejection (major disturbances: F, enthalpy h_F, V, V_n, reflux temperature T_L).
- Manual control of only one composition.

– The transition from manual to automatic control.
– The constraints (on the flow rates and the holdups: levels and pressure).
– Level control.

These factors can be contradictory.

(Waller et al., 1988) studied four control structures with regard to sensitivity:

– The usual scheme of energy balance (L, V) with the levels (condenser and reboiler) controlled by D and B.
– The material balance scheme (D, V) with the levels controlled by L and B.
– Ryskamp scheme $(D/(L+V), V)$ with the levels controlled by $L+D$ and B.
– The two-ratio scheme $(D/(L+V), V/B)$ (Shinskey, 1984) with the levels controlled by $L+D$ and B.

In open loop, the structure (L, V) is the less sensitive and the structure (D, V) the more sensitive to the feed composition disturbances. In closed loop, the quality of structure (L, V) decreases while remaining acceptable. The two-ratio structure gives worse results than predicted by RGA analysis because of the pure delays present in the transfer functions. The Ryskamp scheme that presented a RGA gain near to 1 like the two-ratio scheme gives the best results. The configuration (D, V) is, in fact, equivalent to a compensator obtained by singular value decomposition (Bequette and Edgar, 1988): the strong direction (related to the largest singular value) corresponds to the modification of the external flow rates $(D - B)$, which particularly influences the mean composition $(x_D + x_B)/2$, while the weak direction (related to the lowest singular value) corresponds to the modification of the internal flow rates $(L+V)$, which especially influences the composition difference $(x_D - x_B)$.

According to Skogestad et al. (1990), who studied four configurations: (L, V), (D, V), (D, B), $(L/D, V/B)$ by the frequency RGA method, the best structure for most columns would be the two-ratio configuration $(L/D, V/B)$ in the case of dual control.

The use of a feedforward controller that takes into account the flow rate F disturbance is attractive but difficult to apply, as the high-purity distillation columns are very sensitive to the inventory errors. The combination with a feedback controller is absolutely necessary (Skogestad, 1992).

Note that Levien and Morari (1987) performed internal model control of coupled distillation columns considered as a 3×3 system.

20.4.4 Linear Digital Control

A dynamic identification of the studied column is essential for obtaining the discrete-time linear model of the process. In this domain, few studies are available. Defaye and Caralp (1979) present an example of compared identification on an industrial depropanizer where the outputs are sensitive temperatures and not mole fractions. An observable disturbance is incorporated in the model.

Defaye et al. (1983) used predictive control to control a component concentration on an internal plate of an industrial column. The proposed controller

is, in fact, a dead-beat one.

Sastry et al. (1977) and Dahlqvist (1981) used the basic self-tuning controller developed by Aström and Wittenmark (1973) to control a pilot distillation column. In the case of a single-input single-output process, the model used for on-line identification is

$$A(q^{-1})\, y(t) = b_0\, B(q^{-1})\, u(t-k-1) + \epsilon(t) \tag{20.19}$$

with k as a pure delay, $\epsilon(t)$ a disturbance, and the polynomials

$$\begin{array}{ll} A(q^{-1}) & = 1 + a_1\, q^{-1} + \ldots + a_{n_a}\, q^{-n_a} \\ B(q^{-1}) & = 1 + b_1\, q^{-1} + \ldots + b_{n_b}\, q^{-n_b} \end{array} \tag{20.20}$$

In the single-input single-output case, the self-tuning controller algorithm is performed in two stages: first, estimation and then the calculation of the control law.

Parameters are estimated by the recursive least-squares algorithm (Ljung and Söderström, 1986); first the parameter vector $\theta(t)$ is defined

$$\theta(t) = [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}]^T \tag{20.21}$$

and the observation vector $\phi(t)$

$$\phi(t) = [-y(t), \ldots, -y(t-n_a+1), b_0 u(t-1), \ldots, b_0 u(t-n_b)]^T \tag{20.22}$$

The vector $\theta(t)$ is determined at each time step in order to minimize the criterion

$$S_N(\theta) = \sum_{i=1}^{N} \epsilon_i^2(t) \tag{20.23}$$

by using the following recursive equations

$$\begin{array}{ll} \theta(t) & = \theta(t-1) + K(t)\,[y(t) - b_0\, u(t-k-1) - \phi^T(t-k-1)\,\theta(t-1)] \\ K(t) & = P(t-1)\,\phi(t-k-1)\,[1 + \phi^T(t-k-1)\,P(t-1)\,\phi(t-k-1)]^{-1} \\ P(t) & = \dfrac{1}{\lambda}\left\{P(t-1) - K(t)[1 + \phi^T(t-k-1)\,P(t-1)\,\phi(t-k-1)]K^T(t)\right\} \end{array} \tag{20.24}$$

where λ is the forgetting factor.

The control input is calculated by using the last parameter estimations

$$u(t) = \frac{1}{b_0}[\hat{A}(q^{-1})y(t+1) - y(t+1)] - [\hat{B}(q^{-1})u(t) - u(t)] \; = -\frac{1}{b_0}\phi^T(t)\,\theta(t) \tag{20.25}$$

The algorithm of the self-tuning controller estimates the model parameters in order to obtain good regulation and normally gets close to a minimum variance controller. The least-squares estimations can be biased if the disturbances are correlated. This algorithm demands that some parameters are specified before activating the controller: the controller structure n_a and n_b, the pure delay k,

the sampling period T_e, the scaling factor b_0, the forgetting factor λ, as well as some initial values P_0 and θ_0.

The self-tuning controller can include a feedforward action (Dahlqvist, 1981). To predict the disturbance effect on the process, the process model is modified by including a term for the measured disturbance d according to the equation

$$A(q^{-1})\, y(t) = b_0 B(q^{-1})\, u(t-k-1) + L(q^{-1})\, d(t-k-1) + \epsilon(t) \qquad (20.26)$$

with the parameters of the disturbance term defined by

$$L(q^{-1}) = l_0 + l_1\, q^{-1} + \ldots + l_{n_l}\, q^{-n_l} \qquad (20.27)$$

Sastry et al. (1977) used the self-tuning controller to control only the top composition of a pilot distillation column. Dahlqvist (1981) extended the control to multivariable systems and chose the reboiler vapour flow rate V and the reflux flow rate R as control variables. When the top and bottom compositions are simultaneously controlled, the column must be considered as a multivariable system because of the coupling between the rectification and stripping sections. With two single-input single-output controllers in the simultaneous control, a controller action in a control loop would appear as a disturbance on the other one. The magnitude of the interaction between the loops can be decreased by a convenient choice of the control variables. The algorithm can be extended to treat multivariable systems. The process model can be formulated as the following model with two inputs u and two outputs y

$$\begin{aligned} y^1_{t+k+1} + A^{11}(q^{-1})\, y^1_t + A^{12}(q^{-1})\, y^2_t &= B^1(q^{-1})\, u^1_t + C^1(q^{-1})\, u^2_t \\ y^2_{t+k+1} + A^{22}(q^{-1})\, y^2_t + A^{21}(q^{-1})\, y^1_t &= B^2(q^{-1})\, u^2_t + C^2(q^{-1})\, u^1_t \end{aligned} \qquad (20.28)$$

y_1 and y_2 are the top and bottom compositions, u_1 and u_2 the reboiler vapour and reflux flow rates. The polynomial C is analogous to polynomial A. The parameters of A^{12}, A^{21} and C take into account the internal couplings in the column. The parameters of C^2 are larger than the parameters of C^1, as the coupling between the top composition and the reboiler flow rate is larger than the coupling between the bottom composition and the reflux flow rate.

Woinet et al. (1991) described an adaptive pole-placement control on a pilot distillation column. Note that pole-placement is seldom used as such in adaptive control.

Ohshima et al. (1991) described an industrial application of control of a distillation column of fatty acids presenting important disturbances. A model-based predictive control algorithm was modified to include the prediction of disturbances.

Instead of single-input multi-output open-loop identification, Chou et al. (2000) demonstrate that multi-input multi-output closed-loop identification presents decisive advantages for developing nonlinear models (Wiener models) for control of high-purity distillation columns.

20.4.5 Model Predictive Control

Historically, distillation columns were among the first important processes to have derived profit of advanced control through the development of model pre-

dictive control (MPC) (Richalet et al., 1978). Among the significant advantages of MPC are the multivariable control with many inputs and outputs and the handling of constraints. Since that period of expansion of MPC, many industrial applications have been reported (Qin and Badgwell, 2000, 1996). It must be noted that probably all these applications make use of linear MPC based on the information provided by the responses of the controlled outputs to adequate steps on selected inputs. In this context, the well-known model of the distillation columns is not used except for simulation and nonlinear MPC is not commonly implemented, although it is mentioned (Bock et al., 2000; Diehl et al., 2001). For more details on MPC, refer to Chap. 16.

20.4.6 Bilinear Models

Linear models commonly used in control are obtained by linearization in the neighbourhood of steady state and can describe the dynamic phenomena only in a limited range around this operating point. Nonlinear models are not commonly used in control. Bilinear models (Espana and Landau, 1978) constitute a compromise and allow us to describe the behaviour of columns in a wider domain than simple linear models.

To justify the bilinear model issued from the knowledge model, it is necessary to recall this latter model. The considered column is identical to that described by Fig. 20.1; it has n plates plus the condenser. The simplifying hypotheses are the following:
H1: Liquid-vapour equilibrium is instantaneous, the efficiency is constant on each plate.
H2: The influence of hydrodynamics is negligible.
H3: The vapour holdups are zero.
H4: The condenser is total.
H5: The molar liquid holdups are constant on each plate.
H6: The enthalpies are constant on each plate.
H7: The heat losses are neglected.
H8: The inlet and outlet flows are saturated liquid. The dynamics is described by the classical equations for a plate i ($1 \leq i \leq n+1$):
Mass balance for component j:

$$\frac{M_i\, dx_i^j}{dt} = L_{i+1}\, x_{i+1}^j + V_{i-1}\, y_{i-1}^j - L_i\, x_i^j - V_i\, y_i^j + \delta_{ii_F}\, F\, z_F^j \tag{20.29}$$

Global mass balance (equal to 0 because of H5):

$$\frac{dM_i}{dt} = L_{i+1} + V_{i-1} - L_i - V_i + \delta_{ii_F}\, F \tag{20.30}$$

Energy balance (equal to 0 because of H6 and H7):

$$\frac{M_i\, dh_i}{dt} = L_{i+1}\, h_{i+1} + V_{i-1}\, H_{i-1} - L_i\, h_i - V_i\, H_i + \delta_{ii_F}\, F\, h_F + \delta_{i1}\, \dot{Q}_B \tag{20.31}$$

reboiler subscript 1, condenser subscript $n+1$ (not a real plate in the case of a total condenser).

δ_{ij}: Kronecker symbol (= 1 if $i = j$, = 0 if $i \neq j$)

Three inputs are considered: two control inputs L and $\dot{Q}_B$ (equivalent to V_1) and a disturbance F. The feed composition z_F that would be a disturbance is taken as a parameter. The controlled outputs are x_D and x_B. The system contains $2(n+1)$ linear equations (20.30) and (20.31) to be solved with respect to the $2(n+1)$ flow rates on the plates and the condenser. The internal flow rates L_i and V_i depend linearly on the input vector

$$\boldsymbol{u}^T = [L\,, \dot{Q}_B\,, F] \tag{20.32}$$

The new system can thus be written for a component j according to the vector notation

$$\dot{\boldsymbol{x}}^j = \left[\sum_1^3 B_k^1 u_k\right] \boldsymbol{x}^j + \left[\sum_1^3 B_k^2 u_k\right] \boldsymbol{y}^j(x) + \frac{1}{M_{i_F}}\,\delta_{ii_F}\,F\,z_F^j \tag{20.33}$$

The state Eq. (20.29) is first-order homogeneous with respect to the control vector $\boldsymbol{u}$. The vapour mole fractions y_j are related to the liquid mole fractions x_j by the equilibrium constant k_{ij} that depends on temperature and pressure for a given plate i. If k_{ij} can be considered as constant through the column (at least piecewise), we can write: $y_j = b_j + k_j\,x_j$, and we get the bilinear relation

$$\dot{\boldsymbol{x}}^j = \left[\sum_1^3 B_k u_k\right] \boldsymbol{x}^j + \frac{1}{M_{i_F}}\,\delta_{ii_F}\,F\,z_F^j \tag{20.34}$$

To linearize this expression, it would be necessary to not consider the coupling terms $u_k\,x_j$ between state and control variables.

The expression 20.34 is first-order homogeneous both with respect to the control vector $\boldsymbol{u}$ and the state $\boldsymbol{x}$; this bilinear model thus makes the internal flow rates L_i and V_i independent of the concentrations and linearly dependent on the control vector. It thus corresponds to hypothesis H5.

With the number of state variables being important, it is attractive to perform a reduction of the previous model. In this case, the column is considered to be formed by three compartments (Fig. 20.8), one for the condenser, one for the rectification section and one for the stripping section.

Using the deviation variables with respect to the steady state (exponent e) for state and control variables

$$\delta u_k = u_k - u_k^e \quad \text{and:} \quad \delta x_i = x_i - x_i^e \tag{20.35}$$

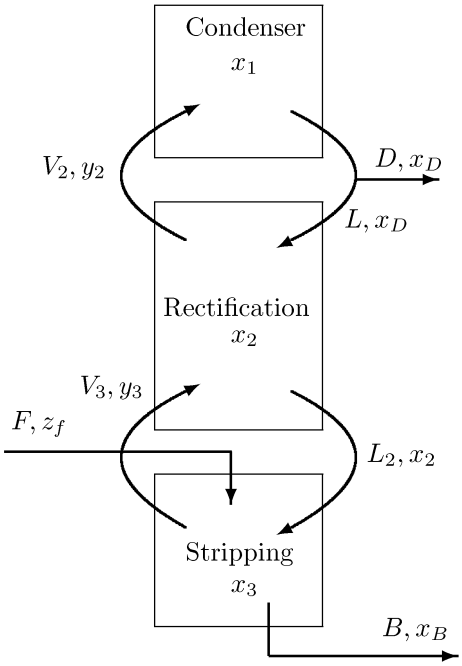


Figure 20.8: Plate aggregation

and the hypothesis of constancy of k_{ij}, a simplified model is obtained

$$
\begin{aligned}
C\,\delta\dot{\boldsymbol{x}} = & \left\{ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} u_1^e + \begin{bmatrix} 0 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} u_2^e + \right. \\
& \left. \begin{bmatrix} -1 & k_2 & 0 \\ 0 & -k_2 & k_3 \\ 0 & 0 & 1-k_3 \end{bmatrix} u_3^e \right\} \delta\boldsymbol{x} \\
& + \left\{ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \delta u_1 + \begin{bmatrix} 0 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \delta u_2 + \right. \\
& \left. \begin{bmatrix} -1 & k_2 & 0 \\ 0 & -k_2 & k_3 \\ 0 & 0 & 1-k_3 \end{bmatrix} \delta u_3 \right\} \delta\boldsymbol{x} \\
& + \begin{bmatrix} 0 \\ 0 \\ z_F - x_3^e \end{bmatrix} \delta u_1 + \begin{bmatrix} 0 \\ x_1^e - x_2^e \\ x_2^e - x_3^e \end{bmatrix} \delta u_2 + \begin{bmatrix} 0 \\ y_1^e - y_2^e \\ y_2^e - y_3^e \end{bmatrix} \delta u_3
\end{aligned}
\tag{20.36}
$$

This equation shows that the global system global includes only eight independent parameters allowing us to describe the dynamic behaviour of the column with three state variables. Equation (20.36) can be written in a more compact

form that is similar to to Eq. (20.34)

$$\delta\dot{\boldsymbol{x}} = \boldsymbol{A}\,\delta\boldsymbol{x} + \left(\sum_{1}^{3} N_k\,\delta u_k\right)\,\delta\boldsymbol{x} + \boldsymbol{B}\,\delta\boldsymbol{u} \tag{20.37}$$

20.4.7 Nonlinear Control

Several nonlinear controls of distillation columns have been published since about 1990. We will particularly cite Lévine and Rouchon (1991); Rouchon (1990) as their control method is actually implemented in several industrial plants, which shows its interest. Rouchon (1990) applied singular perturbation techniques to a knowledge model similar to the previous model given by Eqs. (20.29) to (20.31) to obtain an aggregated nonlinear model usable in control. The objective is to preserve the nonlinear and qualitative properties of the knowledge model (steady-state gains, mole fraction in [0,1], global asymptotic stability) and to calculate a simple and robust nonlinear control law, rejecting the feed composition disturbances.

In this model, the condenser is considered as a true plate (subscript $n+1$). The hypotheses concerning the column are classical and close to those taken for the bilinear model.

In particular, liquid and vapour are perfectly mixed and at equilibrium on each plate: the mass transfer time constants between liquid and vapour are much shorter than the residence time on each plate.

The liquid molar holdup is constant on each plate, the pressure constant and uniform, the vapour molar holdup negligible on each plate: hydrodynamics, pressure and level dynamics are stable and sufficiently fast to be neglected.

In general, the plate geometry and the pressure and level control loops are designed in a way so that these two hypotheses are satisfied for sufficiently smooth inputs L, V, F.

The Lewis hypothesis of constant molar flows in each section of the column is adopted; in the case of components such as hydrocarbons which have close physical properties, it is justified.

While for Espana and Landau (1978) the equilibrium constant k_{ij} was constant, here it is calculated by means of Soave equation of state. Disturbances concern the feed: F, which is measured, and z_F, not measured.

The complete simulation model can be written in the condensed form

$$\dot{x}_i = f_i(x, L, V, F, z_F) \quad , \quad 1 \le i \le n \tag{20.38}$$

for any rectification plate:

$$f_i(x, L, V, F, z_F) = \frac{1}{M_i}\left(L\,x_{i+1} + V\,k(x_{i-1}) - L\,x_i - V\,k(x_i)\right) \tag{20.39}$$

for the feed plate:

$$f_{i_F}(x, L, V, F, z_F) = \frac{1}{M_{i_F}}\begin{pmatrix} L\,x_{i_F+1} + V\,k(x_{i_F-1}) \\ -(L+F)\,x_{i_F} - V\,k(x_{i_F}) + F\,z_F \end{pmatrix} \tag{20.40}$$

and for any stripping plate:

$$f_i(x, L, V, F, z_F) = \frac{1}{M_i}\left((L+F)\,x_{i+1} + V\,k(x_{i-1}) - (L+F)\,x_i - V\,k(x_i)\right) \tag{20.41}$$

The model reduction is performed by considering that it is a two time-scale system, one fast:

$$\epsilon \dot{x}^f = f^f(x^s\,, x^f\,, u\,, w\,, \epsilon) \tag{20.42}$$

the other one slow:

$$\dot{x}^s = f^s(x^s\,, x^f\,, u\,, w\,, \epsilon) \tag{20.43}$$

where (x^s, x^f) is the state vector, u the control vector, w the disturbance vector and ϵ a small positive scalar. It is possible to keep only the slow dynamics corresponding to $\epsilon = 0$. Then, a variable change is realized in order to separate the column into sections of consecutive plates (compartments) and aggregate each section. If a section of m real plates is considered, and if i_a is the subscript of the aggregation plate, it is possible to show that by doing the following change of state variables, which touches only the composition of the aggregation plate

$$x_i = x_i^f \quad \text{for: } 1 \le i \le m\ ,\ i \ne i_a \tag{20.44}$$

$$x_i = x^s = \left(\sum_{j=1}^{m} M_j\,x_j\right)/\bar{M} \quad \text{for: } i = i_a \quad \text{with: } \bar{M} = \sum_{j=1}^{m} M_j \tag{20.45}$$

we obtain a decomposition of the system of equations into a slow sub-system

$$\bar{M}\,\dot{x}^s = L\,x_{m+1} + V\,k(x_0) - L\,x_1^f - V\,k(x_m^f) \quad ,\ i = i_a \tag{20.46}$$

$$0 = L\,x_{i+1} + V\,k(x_{i-1}) - L\,x_i - V\,k(x_i)\ ,\ 1 \le i \le i_a - 1 \text{ and } i_a + 1 \le i \le m \tag{20.47}$$

(where x_{m+1} and $y_0 = k(x_0)$ are the compositions entering the section of m plates), and a fast sub-system

$$\alpha_i\,\bar{M}\,\frac{dx_i}{d\tau} = L\,x_{i+1} + V\,k(x_{i-1}) - L\,x_i - V\,k(x_i)\ ,\ 1 \le i \le i_a - 1 \text{ and } i_a + 1 \le i \le m \tag{20.48}$$

with α and τ defined by

$$M_i = \epsilon\,\alpha_i\,\bar{M} \quad (\text{where } 0 < \epsilon \ll 1) \quad \text{and } \tau = t/\epsilon \tag{20.49}$$

Rouchon (1990) obtains that the dynamics of the section of m plates can be approximated by plates $i \ne i_a$ without holdup and one plate $i = i_a$ having the holdup of the section.

Knowing that the holdups are more important in the condenser and in the reboiler than on the other plates, these two external plates remain unchanged and constitute two compartments. The rest of the column is divided into one to three compartments. In the case of the studied depropanizer, comparisons of open-loop trajectories of product compositions have shown that three compartments for the internal column constitute a good compromise:

one rectification compartment ($i_r \leq i \leq n$) (aggregation plate r)
one feed compartment ($i_r + 1 \leq i \leq i_s - 1$) (aggregation plate i_F)
one stripping compartment ($2 \leq i \leq i_s$) (aggregation plate s)
The entire column is thus represented by a total of 5 compartments.

The complete reduced model is thus as follows:
Equation for the condenser compartment (subscript $n+1$):

$$\bar{M}_{n+1}\,\dot{x}_{n+1} = V\,k(x_n) - V\,x_{n+1} \tag{20.50}$$

Equations for the internal column:

Rectification ($r + 1 \leq i \leq n$):

$$0 = L\,x_{i+1} + V\,k(x_{i-1}) - L\,x_i - V\,k(x_i) \tag{20.51}$$

Rectification aggregation plate:

$$\bar{M}_r\,\dot{x}_r = L\,x_{r+1} + V\,k(x_{r-1}) - L\,x_r - V\,k(x_r) \tag{20.52}$$

Bottom of rectification and top of feed ($i_F + 1 \leq i \leq r - 1$):

$$0 = L\,x_{i+1} + V\,k(x_{i-1}) - L\,x_i - V\,k(x_i) \tag{20.53}$$

Feed aggregation plate:

$$\bar{M}_F\,\dot{x}_{i_F} = L\,x_{i_F+1} + V\,k(x_{i_F-1}) - (L+F)\,x_{i_F} - V\,k(x_{i_F}) + F\,x_F \tag{20.54}$$

Bottom of feed and top of stripping ($s + 1 \leq i \leq i_F - 1$):

$$0 = (L+F)\,x_{i+1} + V\,k(x_{i-1}) - (L+F)\,x_i - V\,k(x_i) \tag{20.55}$$

Stripping aggregation plate:

$$\bar{M}_s\,\dot{x}_s = (L+F)\,x_{s+1} + V\,k(x_{s-1}) - (L+F)\,x_s - V\,k(x_s) \tag{20.56}$$

Bottom of stripping ($2 \leq i \leq s - 1$):

$$0 = (L+F)\,x_{i+1} + V\,k(x_{i-1}) - (L+F)\,x_i - V\,k(x_i) \tag{20.57}$$

Equation for the reboiler (subscript 1)

$$\bar{M}_1\,\dot{x}_1 = (L+F)\,x_2 + (L+F-V)\,x_1 - V\,k(x_1) \tag{20.58}$$

The system of five differential equations has the same tridiagonal structure as the original system and constitutes the control system where L and V are the control variables, x_1 ($= x_B$) and x_{n+1} ($= x_D$) the outputs, F a measured disturbance and z_F an unmeasured disturbance. In fact, it is easier to choose L/V and $(L + F)/V$ as control variables. The system is solved in the same way.

If we denote the two outputs by $y_1 = x_1$ (reboiler) and $y_2 = x_n$ (distillate) solutions of the system

$$\begin{cases} \dfrac{dy_1}{dt} = \phi_1(y_1) \\ \dfrac{dy_2}{dt} = \phi_2(y_2) \end{cases} \tag{20.59}$$

the following condition of PI behaviour resting on the deviation between set point and output is imposed

$$\begin{cases} \phi_1 = & \dfrac{y_1^c(t) - y_1(t)}{\tau_1^P} + \dfrac{\int_0^t (y_1^c(\mu) - y_1(\mu))d\mu}{\tau_1^I \tau_1^P} \\ \phi_2 = & \dfrac{y_2^c(t) - y_2(t)}{\tau_2^P} + \dfrac{\int_0^t (y_2^c(\mu) - y_2(\mu))d\mu}{\tau_2^I \tau_2^P} \end{cases} \tag{20.60}$$

The closed-loop analysis by nonlinear disturbance rejection techniques shows the local stability of feedback control; from the results, it seems that greater stability is likely. The control law depends on (x_B, x_r, x_s, x_D, F); in fact, x_r and x_s are not measured and are estimated by temperature measurements at adequate points of the considered compartments (thermodynamic equilibrium equation $k = f(x)$ and $T = f(x)$) in the feedback control law. The position of the aggregation plates r and s is not very important. Creff (1992) extended the results of Rouchon (1990) to multicomponent columns.

For multicomponent mixtures, it is necessary to measure the temperature not only on a sensitive plate, but also at different points of the column (Yu and Luyben, 1984). Nonlinear observers (Lang and Gilles, 1990) have been developed from the knowledge model of the column; they allow us to estimate the temperature and composition profiles. Knowing that the movement of high mass transfer zones is deciding to characterize the dynamic behaviour of the column, it is possible to place the sensors in an optimal manner. Mejdell and Skogestad (1991) show that it is simply possible to estimate the compositions from temperature measurements and obtains the best results with a partial least-squares static estimator, taking into account uncertainties and different noises.

The nonlinear control law by aggregation allows good disturbance rejection and it is not very sensitive to uncertainties and time delays due to the measurements. This nonlinear control law has been favourably compared with regard to stability and robustness to the linear control law obtained by the geometric approach of Takamatsu et al. (1979); this linear law uses the complete linearized model and is more sensitive to measurement delays (for concentrations obtained by chromatography) than the aggregated nonlinear model. The geometric approach of disturbance rejection (Gauthier et al., 1981) has also been compared (Rouchon, 1990); in this case, the complete nonlinear model is used; it poses the same problem as the work of (Takamatsu et al., 1979) in the presence of delays. The aggregation techniques developed in this chapter in the frame of distillation are general and thus can be applied to a tubular reactor or to any chain that can be modelled as a series of several unit elements. Other

nonlinear studies (Bequette, 1991) have been published, using, for example, nonlinear model predictive control that consists of on-line optimization based on a knowledge model that allows us to numerically calculate the control law while taking into account possible constraints (Eaton and Rawlings, 1990).

20.5 Conclusion

Concerning the much studied physical system of the distillation column, a wide diversity of approaches of the control law is noticed; this diversity is partly related to the use of very different models going from black-box models obtained by identification to knowledge models used in their globality. Through the retained examples, which do not pretend to cover all this area, it can be observed that the more recent approaches are, in general, more complex and time-consuming, as they demand important intellectual investment in the design of the control system. In return, a more stable and safer functioning, in a larger domain around steady state, considering the disturbances, with sometimes the possibility of economic on-line optimization, is obtained. Many problems remain to be solved, in particular, in the case of high-purity distillation of non-ideal multicomponent mixtures. The progress in computing performance allows us to consider solutions that integrate more and more complete knowledge models. However, the handling of constraints, the robustness related to the modelling and measurement errors, to the measurement delays, will always be of great importance.

References

J. Albet, J.M. LeLann, X. Joulia, and B. Khoeret. Rigorous simulation of multicomponent multisequence batch reactive distillation. In L. Puigjaner and A. Espuna, editors, *Computer-Oriented Process Engineering*, Amsterdam, 1991. Elsevier.

Y. Arkun and C.O. Morgan. On the use of the structured singular value for robustness analysis of distillation column control. *Comp. Chem. Engng.*, 12 (4):303–306, 1988.

Y. Arkun, B. Manousiouthakis, and A. Palazoglu. Robustness analysis of process control systems. A case study of decoupling control in distillation. *Ind. Eng. Chem. Process Des. Dev.*, 23:93–101, 1984.

K.J. Aström and B. Wittenmark. On self tuning regulators. *Automatica*, pages 185–199, 1973.

B.W. Bequette. Nonlinear control of chemical processes: a review. *Ind. Eng. Chem. Res.*, 30:1391–1413, 1991.

B.W. Bequette and T.F. Edgar. A dynamic interaction index based on set-point transmittance. *AIChE J.*, 34:849–852, 1988.

H. G. Bock, M. Diehl, J.P. Schlöder, F. Allgöwer, R. Findeisen, and Z. Nagy. Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. In *International Symposium on Advanced Control of Chemical Processes*, pages 695–702, Pisa Italy, 2000.

I.T. Cameron, C.A. Ruiz, and R. Gani. A generalized model for distillation columns - II Numerical and computational aspects. *Comp. Chem. Engng.*, 10(3):199–211, 1986.

C.T. Chou, H.H.J. Bloemen, V. Verdult, T.T.J. Van Den Boom, T. Backx, and M. Verhaegen. Nonlinear identification of high-purity distillation columns. In *IFAC SYSID, Symposium on System Identification*, Santa Barbara, 2000.

Y. Creff. *Sur la Dynamique et la Commande des Colonnes Multicomposées.* Phd thesis, Ecole des Mines de Paris, Paris, 1992.

S.A. Dahlqvist. Control of a distillation column using self-tuning regulators. *Can. J. Chem. Eng.*, 59(2):118–127, 1981.

G. Defaye and L. Caralp. Essais de modélisation d'une distillation industrielle par identification dynamique. Etude d'un dépropaniseur. *Chem. Eng. J.*, 18: 131–141, 1979.

G. Defaye, L. Caralp, and P. Jouve. A simple deterministic predictive control algorithm and its application to an industrial chemical process: a distillation column. *Chem. Eng. J.*, 27:161–166, 1983.

M. Diehl, I. Uslu, R. Findeisen, S. Schwartzkopf, F. Allgöwer, H.G. Bock andT. Bürner, E.D. Gilles, A. Kienle, J.P. Schlöder, and E. Stein. Real-time optimization for large scale processes: nonlinear model predictive control of a high-purity distillation column. In Groeschel, Krumke, and Rambau, editors, *Online Optimization of Large Scale Systems: State of the Art*, pages 363–383. Springer, 2001.

J.W. Eaton and J.B. Rawlings. Feedback control of chemical processes using on-line optimization techniques. *Comp. Chem. Engng.*, 14:469–479, 1990.

M. Espana and I.D. Landau. Reduced order bilinear models for distillation columns. *Automatica*, 14:345–355, 1978.

S.E. Gallun and C.D. Holland. Gear's procedure for the simultaneous solution of differential and algebraic equations with application to unsteady-state distillation problems. *Comp. Chem. Engng.*, 6:231–244, 1982.

R. Gani, C.A. Ruiz, and I.T. Cameron. A generalized model for distillation columns - I Model description and applications. *Comp. Chem. Engng.*, 10 (3):181–198, 1986.

J.P. Gauthier, G. Bornard, S. Bacha, and M. Idir. Rejet de perturbations pour un modèle nonlinéaire de colonne à distiller. In *Outils et Modèles Mathématiques pour l'Automatique, l'Analyse des Systèmes et le Traitement du Signal*, pages 659–673. Editions du CNRS, 1981.

C.W. Gear and L.R. Petzold. Ode methods for the solution of differential/algebraic systems. *SIAM J. Numer. Anal.*, 21(4):716–728, 1984.

A.C. Hindmarsh and L.R. Petzold. Algorithms and software for ordinary differential equations and differential-algebraic equations, part II: Higher-order methods and software packages. *Computers in Physics*, 9(2):148, 1995.

C.D. Holland. *Fundamentals of Multicomponent Distillation.* McGraw-Hill, New York, 1981.

M. Hovd and S. Skogestad. Simple frequency-dependent tools for control system analysis, structure selection and design. *Automatica*, 28(5):989–996, 1992.

E.W. Jacobsen and S. Skogestad. Instability of distillation columns. *AIChE J.*, 40(9):1466–1478, 1994.

L. Lang and E.D. Gilles. Nonlinear observers for distillation columns. *Comp. Chem. Eng.*, 14(11):1297–1301, 1990.

P. Lang, H. Yatim, P. Moszkowicz, and M. Otterbein. Batch extractive distillation under constant reflux ratio. *Comp. Chem. Engng.*, 18(11/12):1057–1069, 1995.

K.L. Levien and M. Morari. Internal model control of coupled distillation columns. *AIChE J.*, 33(1):83–98, 1987.

J. Lévine and P. Rouchon. Quality control of binary distillation columns based on nonlinear aggregated models. *Automatica*, 27(3):463–480, 1991.

L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, Massachussetts, 1986.

W.L. Luyben. Distillation decoupling. *AIChE J.*, (3):198–203, 1970.

T. Mejdell and S. Skogestad. Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression. *Ind. Eng. Chem. Res.*, 30:2543–2555, 1991.

A. Niederlinski. Two-variable distillation control: Decouple or not decouple. *AIChE J.*, 17(5):1261–1263, 1971.

M. Ohshima, H. Ohno, I. Hashimoto, M. Sasajima, M. Maejima, K. Tsuto, and T. Ogawa. Model predictive control with adaptive disturbance prediction and its application to fatty acid distillation columns control. AIChE Annual Meeting, Los Angeles, 1991.

J.M. Prausnitz and P.L. Chueh. *Computer Calculations for High-Pressure Vapor Liquid Equilibria.* Prentice Hall, Englewood Cliffs, 1968.

J.M. Prausnitz, C.A. Eckert, R.V. Orye, and J.P. O'Connel. *Computer Calculations for Multicomponent Vapor Liquid Equilibria.* Prentice Hall, Englewood Cliffs, 1967.

S.J. Qin and T.A. Badgwell. An overview of nonlinear model predictive control applications. In F. Allgöwer and A. Zheng, editors, *Non Linear Model Predictive Control*, pages 369–392. Birkhäuser, Basel, 2000.

S.J. Qin and T.A. Badgwell. An overview of industrial model control technology. In *Chemical Process Control - CPC V*, pages 232–255, Tahoe, California, 1996.

J. Richalet, A. Rault, J.L. Testud, and J. Papon. Model predictive heuristic control: Applications to industrial processes. *Automatica*, 14:413–428, 1978.

L.M. Rose. *Distillation Design in Practice*. Elsevier, Amsterdam, 1985.

P. Rouchon. *Simulation Dynamique et Commande Non Linéaire des Colonnes à Distiller*. Phd thesis, Ecole des Mines de Paris, Paris, 1990.

C.A. Ruiz, I.T. Cameron, and R. Gani. A generalized model for distillation columns - III Study of startup operations. *Comp. Chem. Engng.*, 12(1):1–14, 1988.

V.A. Sastry, D.E. Seborg, and R.K. Wood. Self-tuning regulator applied to a binary distillation column. *Automatica*, 13:417–424, 1977.

F.G. Shinskey. *Distillation Control*. McGraw-Hill, New York, 1984.

S. Skogestad. Dynamics and control of distillation columns - A critical survey. IFAC-Symposium Dycord+'92 Maryland, 1992.

S. Skogestad. Dynamics and control of distillation columns. *Trans. IChemE*, 75:539–562, 1997.

S. Skogestad and P. Lundstrom. Mu-optimal LV-control of distillation columns. *Comp. Chem. Engng.*, 14(4/5):401–413, 1990.

S. Skogestad and M. Morari. Control configuration selection for distillation columns. *AIChE J.*, 33(10):1620–1635, 1987a.

S. Skogestad and M. Morari. Implications of large RGA elements on control performance. *Ind. Eng. Chem. Res.*, 26:2323–2330, 1987b.

S. Skogestad and M. Morari. Understanding the dynamic behavior of distillation columns. *Ind. Eng. Chem. Res.*, 27:1848–1862, 1988.

S. Skogestad, M. Morari, and J.C. Doyle. Robust control of ill-conditioned plants : High-purity-distillation. *IEEE Trans. Auto. Control*, 33:1092–1095, 1988.

S. Skogestad, P. Lundstrom, and E.W. Jacobsen. Selecting the best distillation control configuration. *AIChE J.*, 36(5):753–764, 1990.

T. Takamatsu, I. Hashimoto, and Y. Nakai. A geometric approach to multivariable control system design of a distillation column. *Automatica*, 15: 387–402, 1979.

J. Unger, A. Kröner, and W. Marquardt. Structural analysis of differential-algebraic equation systems - Theory and applications. *Comp. Chem. Engng.*, 19(8):867–882, 1995.

K.V. Waller, K.E. Häggblom, P.M. Sandelin, and D. H. Finnerman. Disturbance sensitivity of distillation control structures. *AIChE J.*, 34(5):853–858, 1988.

F. Weischedel and T.J. McAvoy. Feasibility of decoupling in conventionally controlled distillation columns. *Ind. Eng. Chem. Fundam.*, 19:379–384, 1980.

R. Woinet, G. Thomas, and J. Bordet. Adaptive control based on pole placement: An experimental test on a binary distillation column. *Chem. Eng. Sci.*, 46(4):949–957, 1991.

R.K. Wood and M.W. Berry. Terminal composition control of a binary distillation column. *Chem. Eng. Sci.*, 28:1707–1717, 1973.

C.C. Yu and W.L. Luyben. Use of multiple temperatures for the control of multicomponent distillation columns. *Ind. Eng. Chem. Process Des. Dev.*, 23:590–597, 1984.

Chapter 21

Examples and Benchmarks of Typical Processes

In this chapter, a series of different processes that are used as examples or benchmarks in the literature to test control strategies are given.

21.1 Single-Input Single-Output Processes

21.1.1 Description by Transfer Functions

Process with Large Time Constant and Small Delay

The following plant continuous transfer function

$$G(s) = \frac{100}{100s + 1} \exp(-s) \tag{21.1}$$

poses some problems, in particular, in discrete-time identification for further use of a discrete-time controller. A high closed-loop bandwidth ω_b is desired. Using the common following rule for the sampling period T_s

$$T_s \leq \frac{2\pi}{\omega_b} \tag{21.2}$$

Lunström et al. (1995) mention that to obtain a truncation error lower than 5%, 300 step coefficients are needed. However, this introduces a conflict with industrial practice, where a lower number of coefficients is retained: 30 according to Cutler and Ramaker (1980).

To demonstrate the influence of truncation, use a given discrete-time controller and analyse the response to a step input occurring at time $t = 10$. Lunström et al. (1995) show that a DMC controller leads to closed-loop instability due to the large model error.

21.1.2 Description by a State-Space Knowledge Model

Van de Vusse Reactor

The Van de Vusse reaction is operated in a continuous stirred tank reactor. It is often used in control literature (Chikkula and Lee, 2000; Stack and Doyle III, 1997). The reaction is

$$\begin{array}{l} \mathrm{A} \rightarrow \mathrm{B} \rightarrow \mathrm{C} \\ 2\,\mathrm{A} \rightarrow \mathrm{D} \end{array} \tag{21.3}$$

with the following material balances for components A and B

$$\begin{array}{l} \dfrac{dC_A}{dt} = -k_1\,C_A - k_3\,C_A^2 + \dfrac{F}{V}\,(C_{Af} - C_A) \\ \dfrac{dC_B}{dt} = k_1\,C_A - k_2\,C_B - \dfrac{F}{V}\,C_B \end{array} \tag{21.4}$$

The nominal operating conditions are given in Table 21.1. The control problem consists of regulating the concentration of intermediate product C_B by manipulating the inlet flow rate F. F_0 is the steady-state flow rate corresponding to a steady-state value of $C_B = 1$ mol.l^{-1}. When F increases from 0 to 250 l.h^{-1}, the steady-state concentration C_B first strongly increases, then decreases. Thus, the process is highly nonlinear with a change of sign of the steady-state gain. Also, when the steady-state concentration C_B is lower than the maximum steady-state concentration C_B, the reactor presents an inverse response, thus is nonminimum-phase. This corresponds to the proposed operating point of Table 21.1.

Table 21.1: Nominal variables and main parameters of the Van de Vusse CSTR

Reactor volume	$V = 1$ l
Nominal feed flow rate	$F_0 = 25$ l.h^{-1}
Feed concentration of reactant A	$C_{Af} = 10$ mol.l^{-1}
Kinetic constant	$k_1 = 50$ h^{-1}
Kinetic constant	$k_2 = 100$ h^{-1}
Kinetic constant	$k_3 = 10$ l.mol^{-1}.h^{-1}

Stability Study of a Continuous Stirred Tank Reactor

Uppal et al. (1974) have proposed a parametric model for a continuous stirred tank reactor which allowed them to study the occurrence of multiple steady

states. This reactor is modelled by the following set of equations

$$\begin{aligned} V\frac{dC_A}{dt'} = &\quad \lambda\, F\, C_{Af} + F\,(1-\lambda)\, C_A - F\, C_A - V\, k_0\, \exp(-\frac{E}{RT})\, C_A \\ V\,\rho\, C_p \frac{dT}{dt'} = &\quad \rho\, C_p\, F\,(\lambda\, T_f + (1-\lambda)\, T - T) + V\,(-\Delta H)\, k_0\, \exp(-\frac{E}{RT})\, C_A \\ &\quad -h\, A\,(T - T_c) \end{aligned} \tag{21.5}$$

$(1-\lambda)$ corresponds to the fraction of recycled outlet stream.

In order to obtain a system in dimensionless form

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2) \\ \dot{x}_2 &= f_2(x_1, x_2) \end{aligned} \tag{21.6}$$

the following auxiliary variables are introduced

$$\begin{aligned} x_1 &= \frac{C_{Af} - C_A}{C_{Af}} & x_2 &= \frac{T - T_f}{T_f}\left(\frac{E}{RT_f}\right) \\ \tau &= \frac{V}{F\,\lambda} & t &= \frac{t'\,F\,\lambda}{V} = \frac{t'}{\tau} \\ Da &= \frac{k_0\, \exp(-\gamma)\, V}{F\,\lambda} = k_0\, \exp(-\gamma)\, \tau & B &= \frac{(-\Delta H)\, C_{Af}}{\rho\, C_p\, T_f}\left(\frac{E}{RT_f}\right) \\ x_{2c} &= \frac{T_c - T_f}{T_f}\left(\frac{E}{RT_f}\right) & \beta &= \frac{h\, A}{F\,\lambda\,\rho\, C_p} = \frac{h\, A\,\tau}{V\,\rho\, C_p} \\ \gamma &= \frac{E}{RT_f} \end{aligned} \tag{21.7}$$

At this stage, a stability study can be performed.

Show that the stability condition for a dynamic system of dimension n modelled as

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\,\boldsymbol{x} \tag{21.8}$$

results for a system of dimension 2 in the following condition

$$\det(\boldsymbol{A}) > 0 \qquad ; \quad \text{trace of } (\boldsymbol{A}) < 0 \tag{21.9}$$

Show that the stationary solution (x_1^s, x_2^s) of the system (21.6) satisfies the equations

$$\begin{aligned} x_2^s &= \frac{B\, x_1^s + \beta\, x_{2c}}{1+\beta} \\ Da &= \frac{x_1^s}{1 - x_1^s}\, \exp\left(\frac{-(B\, x_1^s + \beta\, x_{2c})}{(1+\beta) + \dfrac{1}{\gamma}(B\, x_1^s + \beta\, x_{2c})}\right) \end{aligned} \tag{21.10}$$

Plot the curve representing the dependence of x_1^s (ordinate) with respect to Da (abscissa) for given values of B, β, γ, x_{2c} (choose $B = 14.94$, $\beta = 2.09$, $\gamma = 25$, $x_{2c} = -2.78$). Interpret this curve with respect to stability.

From the values of Table 21.2, determine whether the operating point is stable or unstable.

An Unstable Continuous Stirred Tank Reactor

Consider again the unstable continuous stirred tank reactor studied by Uppal et al. (1974). Its behaviour is interesting from a control viewpoint as it is highly nonlinear and open-loop unstable. It has been used by Henson and Seborg (1993).

The reaction is first-order A $\rightarrow$ B. The model is given as

$$\begin{aligned} \frac{dC_A}{dt} &= \frac{q}{V}(C_{Af} - C_A) - k_0 \exp\left(-\frac{E}{RT}\right) C_A \\ \frac{dT}{dt} &= \frac{q}{V}(T_f - T) - \frac{\Delta H}{\rho C_p} k_0 \exp\left(-\frac{E}{RT}\right) C_A + \frac{UA}{V\rho C_p}(T_c - T) \end{aligned} \tag{21.11}$$

The manipulated variable is the coolant temperature T_c and the controlled variable is the reactor temperature T. C_A is the effluent concentration. C_{Af}, T_f and q are respectively the feed concentration, temperature and flow rate. Table 21.2 gives the values of the nominal variables and main parameters.

Table 21.2: Nominal variables and main parameters of CSTR defined in (Henson and Seborg, 1993)

Feed flow rate	$q = 100$ l.min^{-1}
Feed concentration of reactant A	$C_{Af} = 1$ mol.l^{-1}
Feed temperature	$T_f = 350$ K.
Reactor volume	$V = 100$ l
Heat transfer coefficient	$UA = 5 \times 10^4$ J.min^{-1}.K^{-1}
Preexponential factor	$k_0 = 7.2 \times 10^{10}$ min^{-1}
Reduced activation energy	$E/R = 8750$ K
Heat of reaction	$-\Delta H = 5 \times 10^4$ J.mol^{-1}
Density of reactor contents	$\rho = 1000$ g.l^{-1}
Heat capacity of reactor contents	$C_p = 0.239$ J.g^{-1}.K^{-1}
Coolant temperature	$T_c = 311.1$ K
Concentration of reactor contents	$C_A = 9.3413 \times 10^{-2}$ mol.l^{-1}
Reactor temperature	$T = 385$ K

Henson and Seborg (1993) show a severe nonlinear behaviour by comparing the open-loop responses of C_A to ± 5 step changes in T_c. In this reactor, they have applied different control strategies. In particular, they exactly discretized the previous continuous model with a sampling period $T_s = 0.1$ min and obtained the discrete-time transfer function

$$\tilde{G}(z) = z^{-1} \frac{0.2801 - 0.0853 z^{-1}}{1 - 1.592 z^{-1} - 0.7801 z^{-2}} \tag{21.12}$$

They then compared the performances of internal model control (IMC) and of nonlinear IMC. For IMC, they used a first-order filter with constant $\alpha = 0.368$,

corresponding to a continuous filter of time constant 0.1 min. The performances were compared for ±10 K step changes in the set point and for −25% unmeasured step disturbance in the feed flow rate q.

A Chemical Reactor

Let us take again the chemical reactor modelled by equations (19.9) in Chap. 19. Consider only the thermal behaviour of the reactor, in the absence of chemical reaction, expressed by the two equations of energy balance leading to the derivatives of the reactor and jacket temperatures

$$\begin{aligned} m\,C_p\,\frac{dT}{dt} &= F_o\,\rho\,C_p(T_o - T) - UA\,(T - T_j) \\ V_j\,\rho_j\,C_{pj}\,\frac{dT_j}{dt} &= F_j\,\rho_j\,C_{pj}(T_{j,in} - T_j) + UA\,(T - T_j) \end{aligned} \tag{21.13}$$

as well as the linear dependency of the inlet temperature $T_{j,in}$ in the jacket with respect to the position α of the three-way valve according to

$$T_{j,in} = \alpha\,T_c + (1 - \alpha)\,T_f \tag{21.14}$$

Then, the expressions of the Laplace transforms of the temperature T of the reactor contents and the jacket T_j can be obtained. The manipulated variable is either $T_{j,in}$ or α according to the desired degree of subtlety. From a block diagram for the reactor temperature control, the relations between T and T_j can be emphasized. Lastly, different types of control from PID to more efficient ones can be performed.

A Simplified Heat Exchanger

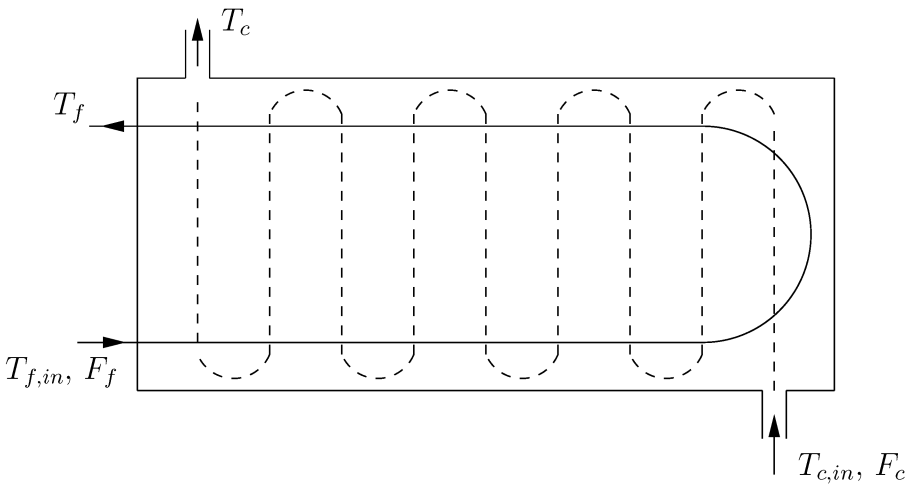


Figure 21.1: Heat exchanger

Consider the heat exchanger represented by Fig. 21.1. We wish to control temperature T_c of the cold fluid (subscript "c") by manipulating the mass flow rate F_h of hot fluid (subscript "h"). The temperatures $T_{h,in}$ (inlet of hot fluid) and $T_{c,in}$ (inlet of cold fluid) are likely to change as well as the mass flow rate of cold fluid F_c. In order to simplify the study, it is assumed that the outlet temperature T_h of the hot fluid is also the same at any point of the heat exchanger (thus considered from this point of view as a perfectly stirred reactor). The same hypotheses are taken for the outlet temperature T_c of the cold fluid. A global heat transfer coefficient h is considered between the tubes (for the cold fluid) and the shell (for the hot fluid). The steady-state data are given in Table 21.3.

Table 21.3: Steady-state values of the different variables for the heat exchanger

Variable	Symbol	Value
Hot flow rate	F_h	29.68 kg.s^{-1}
Cold flow rate	F_c	37.32 kg.s^{-1}
Hot temperature at the inlet	$T_{h,in}$	420 K
Hot temperature at the outlet	$T_{h,out}$	380 K
Cold temperature at the inlet	$T_{c,in}$	295 K
Cold temperature at the outlet	$T_{c,out}$	330 K
Density of the hot fluid	ρ_h	900 kg.m^{-3}
Density of the cold fluid	ρ_c	800 kg.m^{-3}
Mass of hot fluid in the exchanger	m_h	180 kg
Mass of cold fluid in the exchanger	m_c	640 kg
Heat capacity of the hot fluid	C_{ph}	2200 J.kg^{-1}.K^{-1}
Heat capacity of the cold fluid	C_{pc}	2000 J.kg^{-1}.K^{-1}
Global heat transfer coefficient	h	740 W.m^{-2}.K^{-1}
Exchange surface	A	70.6 m^2

Show that the analytical equations of the balances of this heat exchanger lead to the following linearized model (21.15) given in numerical form

$$\begin{aligned} \bar{T}_c &= \frac{-0.5517}{10.09\,s+1}\,\bar{F}_c + \frac{0.5883}{10.09\,s+1}\,\bar{T}_{c,in} + \frac{0.4117}{10.09\,s+1}\,\bar{T}_h \\ \bar{T}_h &= \frac{0.7487}{3.369\,s+1}\,\bar{F}_h + \frac{0.5555}{3.369\,s+1}\,\bar{T}_{h,in} + \frac{0.4445}{3.369\,s+1}\,\bar{T}_c \end{aligned} \tag{21.15}$$

Considering the role played by the temperatures T_h and T_c, the equations of this heat exchanger present a great analogy with the jacketed chemical reactor (Eqs. 21.13) considered with respect to the temperatures T of the reactor and T_j of the jacket. It is interesting to demonstrate this analogy by a block diagram including the control elements.

Assume that an actuator is used with its transfer function being a pure gain $K_a = 0.1$ and a sensor whose transfer function is also a pure gain $K_m = 0.02$.

On the other hand, the sensor is placed in the outlet pipe of the cold fluid which introduces a delay equal to 4 s. The control of this heat exchanger can be performed in many ways.

A Fed-Batch Bioreactor

Fed-batch bioreactors offer numerous possibilities of study in dynamic optimization, state estimation and control. Thus Rodrigues and Filho (1996) studied a fed-batch bioreactor of penicillin production and tested it with predictive DMC control. The bioreactor cited hereafter is used by Srinivasan et al. (2002a) to demonstrate the maximization of the concentration of product P (also penicillin) for a given final time. X is biomass, S is substrate and F is the manipulated feed flow rate. The model equations are

$$\begin{aligned}
\dot{X} &= \mu(S)\, X - \frac{F}{V} X \\
\dot{S} &= -\frac{\mu(S)\, X}{Y_X} - \frac{\nu X}{Y_P} + \frac{F}{V}\,(S_{in} - S) \\
\dot{P} &= \nu\, X - \frac{F}{V} P \\
\dot{V} &= F
\end{aligned} \tag{21.16}$$

with

$$\mu(S) = \frac{\mu_m\, S}{K_m + S + S^2/K_i} \tag{21.17}$$

and the model parameters, the initial values and the operating constraints are given in Table 21.4.

Table 21.4: Model parameters, initial values and operating constraints for the fed-batch bioreactor cited by (Srinivasan et al., 2002a)

Constant	$\mu_m = 0.02$ l.h^{-1}
Constant	$K_m = 0.05$ g.l^{-1}
Inhibition constant	$K_i = 5$ g.l^{-1}
Yield	$Y_X = 0.5$ g$[X]$.g$^{-1}[S]$
Yield	$Y_P = 1.2$ g$[P]$.g$^{-1}[S]$
Parameter	$\nu = 0.004$ l.h^{-1}
Initial biomass concentration	$X_0 = 1$ g/l
Initial substrate concentration	$S_0 = 0.5$ g/l
Initial product concentration	$P_0 = 0$ g/l
Initial volume of the reactor	$V_0 = 150$ l
Substrate concentration in the feed	$S_{in} = 200$ g/l
Final time	$t_f = 150$ h
Constraints on the feed flow rate	$0 \leq F \leq 1$ l.h^{-1}
Constraints on the biomass concentration	$X \leq 3.7$ l.h^{-1}

Srinivasan et al. (2002b) treated the same example, but with different parameters and also different operating conditions. They also provide several examples of various processes in view of dynamic optimization.

Another model for batch penicillin production, including the influence of pH and temperature, is given by Birol et al. (2002).

21.1.3 Description by a Linear State-Space Model

A Continuous Stirred Tank Reactor

This reactor (Ramirez, 1994) has been used (Choi et al., 2000) to demonstrate constrained linear quadratic optimal control. The concerned reaction is the decomposition of hydrogen peroxide H_2O_2 into liquid water and gaseous oxygen. Temperature and hydrogen peroxide concentration in the reactor are the two states. Temperature is directly measured and hydrogen peroxide concentration can be estimated from the reaction rate by monitoring gaseous oxygen evolution rate. The manipulated input is a cooling valve control signal. This input is constrained in the range $[-1, 1]$.

The linearized continuous state-space model is

$$\dot{\boldsymbol{x}}(t) = \begin{bmatrix} -0.0556 & -0.05877 \\ 0.01503 & -0.005887 \end{bmatrix} \boldsymbol{x}(t) + \begin{bmatrix} 0 \\ -0.03384 \end{bmatrix} u(t) \tag{21.18}$$

or the following discrete state-space model is

$$\boldsymbol{x}(k+1) = \begin{bmatrix} 0.8347 & -0.1811 \\ 0.0463 & 0.9754 \end{bmatrix} \boldsymbol{x}(k) + \begin{bmatrix} 0.0108 \\ -0.1137 \end{bmatrix} u(k) \tag{21.19}$$

21.2 Multivariable Processes

21.2.1 Matrices of Continuous Transfer Functions

A Subsystem of a Heavy Oil Fractionator

This process is a top 2×2 subsystem of the heavy oil fractionator modelled in the Shell Standard Control (Prett and Garcia, 1988). It has been used by Zafiriou (1990) and Vuthandam et al. (1995). The process transfer function matrix is

$$\bar{G}(s) = \begin{bmatrix} \dfrac{4.05\,\exp(-27s)}{50s+1} & \dfrac{1.77\,\exp(-28s)}{60s+1} \\ \dfrac{5.39\,\exp(-18s)}{50s+1} & \dfrac{5.72\,\exp(-14s)}{60s+1} \end{bmatrix} \tag{21.20}$$

where the time unit is minutes. Vuthandam et al. (1995) used a sampling interval of 4 min and the following constraints

$$\begin{aligned} -3.0 &\le \Delta u_2(k) \le 3.0 \\ -5.0 &\le u_2(k) \le 5.0 \\ -0.5 &\le y_1(k+7) \le 0.5 \end{aligned}$$

with set points: $y_1^s = y_2^s = 0.0$ and disturbances: $d_1 = 1.2$ and $d_2 = -0.5$.

A Three-product Distillation Column

This example of a three-product distillation column (Deshpande, 1989) has been used (Al-Ghazzawi et al., 2001) for testing model predictive control. This column separates an ethanol-water feed into ethanol at two locations, overhead and sidestream, and water at the bottom.

The dynamic model is

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \dfrac{0.66\exp(-2.6s)}{6.7s+1} & \dfrac{-0.61\exp(-3.5s)}{8.64s+1} & \dfrac{-0.0049\exp(-s)}{9.06s+1} \\ \dfrac{1.11\exp(-6.5s)}{3.25s+1} & \dfrac{-2.36\exp(-3s)}{5s+1} & \dfrac{-0.012\exp(-1.2s)}{7.09s+1} \\ \dfrac{-34\exp(-9.2s)}{8.15s+1} & \dfrac{46.2\exp(-9.4s)}{10.9s+1} & \dfrac{(10.1s+0.87)\exp(-s)}{73.13s^2+22.7s+1} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$
$$+ \begin{bmatrix} \dfrac{0.14\exp(-12s)}{6.2s+1} & \dfrac{(-0.029s-0.011)\exp(-2.66s)}{114.85s^2+22.84s+1} \\ \dfrac{0.53}{6.9s+1}\exp(-10.5s) & \dfrac{(-0.0627s-0.0032)\exp(-2.66s)}{65.17s^2+16.23s+1} \\ \dfrac{-11.54\exp(-0.6s)}{7.01s+1} & 0 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \tag{21.21}$$

y_1 is the ethanol overhead mole fraction, y_2 is the ethanol sidestream mole fraction and y_3 is the temperature at tray number 19 (°C). u_1 is the overhead reflux flow rate ($\mathrm{l.s^{-1}}$), u_2 is the sidestream draw-off flow rate ($\mathrm{l.s^{-1}}$), u_3 is the reboiler steam pressure (kPa). d_1 and d_2 are respectively the feed flow rate ($\mathrm{l.s^{-1}}$) and feed temperature disturbances (°C). A typical sampling time $T_s = 1\,\mathrm{s}$ can be used. However, this is likely to pose problems due to the large time delays.

A Lime Kiln

The lime kiln is a continuous endothermic reactor in a pulp mill where calcium carbonate is transformed into calcium oxide by means of the heat produced by a flame at the end of the kiln bed. Because it is one of the most energy-consuming processes, its control is particularly interesting. The model is given by Zanovello and Budman (1999), who used a variant of model predictive control for this

system. The transfer matrix is

$$\begin{bmatrix} T_{\text{hot}} \\ T_{\text{cold}} \\ O_2 \end{bmatrix} = \begin{bmatrix} \dfrac{80\exp(-2s)}{152.5s+1} & \dfrac{1.1\exp(-15s)}{(40s+1)(50s+1)} \\ \dfrac{-13.65\exp(-2s)}{6.5s+1} & \dfrac{0.3\exp(-2s)}{100s+1} \\ -0.72 & -0.0075 \end{bmatrix} \begin{bmatrix} u_{\text{air}} \\ u_{\text{gas}} \end{bmatrix} + \begin{bmatrix} \dfrac{0.65\exp(-149s)}{25s+1} \\ \dfrac{-0.785\exp(-25s)}{82.5s+1} \\ \dfrac{-0.011\exp(-25s)}{33s+1} \end{bmatrix} \begin{bmatrix} d_{\text{feed}} \end{bmatrix} \tag{21.22}$$

The front-end temperature T_{hot} must be sufficiently high to ensure complete conversion, but not excessively to avoid damage to the refractory lining. The cold-end temperature T_{cold} must not be too low, to avoid agglomeration of the feed, and not too high, to avoid damage to the inner mechanical system of chains. The oxygen concentration O_2 in the exit gases is an indicator of correct fan power if excessive and also of H_2 presence if low. The manipulated variables are the air and natural gas flow rates, respectively u_{air} and u_{gas}. The disturbance is the feed flow rate d_{feed}.

Nominal operating conditions are: $T_{\text{hot}} = 1000\,^{\circ}\text{C}$, $T_{\text{cold}} = 200\,^{\circ}\text{C}$, $0.6 < O_2 < 5\%$, $u_{\text{air}} = 3$mm Hg absolute pressure (approximately 2500 ton/day of air), $u_{\text{gas}} = 1100\ \text{m}^{-3}\text{h}^{-1}$, MUD feed = 800 ton/day, percentage of CaO recovery = 90%.

A robustness study can be performed by assuming the uncertainty tabulated in Table 21.5 in the parameters.

Table 21.5: Parameter uncertainty for lime kiln (Zanovello and Budman, 1999)

$\frac{y}{u}$	Gain K	Delay t_d
$T_{\text{hot}}/u_{\text{air}}$	$75 \le K \le 85$	$1 \le t_d \le 3$
$T_{\text{hot}}/u_{\text{gas}}$	$1 \le K \le 1.2$	$13 \le t_d \le 17$
$T_{\text{cold}}/u_{\text{air}}$	$-15 \le K \le -12$	$0.1 \le t_d \le 0.3$
$T_{\text{cold}}/u_{\text{gas}}$	$0.2 \le K \le 0.4$	$1 \le t_d \le 3$
O_2/u_{air}	$-0.74 \le K \le -0.7$	
O_2/u_{gas}	$-0.008 \le K \le -0.007$	

A typical sampling time is $T_s = 5$min. The tracking can be studied for a set point change of $10\,^{\circ}\text{C}$ in T_{cold}. The disturbance rejection can be tested for a step disturbance of 80 ton/day in MUD feed.

The following cost function can be defined

$$J = \begin{array}{l} \|\Gamma[\hat{\boldsymbol{y}}(k+1|k) - \boldsymbol{y}^{ref}(k+1)]\| + \|\Lambda\Delta\boldsymbol{u}(k)\|^2 + \\ \|w_j[\hat{\boldsymbol{y}}(k+1|k) - \boldsymbol{y}^{ctr}(k+1)\| \end{array} \tag{21.23}$$

The previous cost function is different from the one used by (Zanovello and Budman, 1999) in the penalty term. Γ and Λ are diagonal weight matrices, $\boldsymbol{y}^{ref}$ is the reference trajectory, $\boldsymbol{y}^{ctr}$ is the constraint.

The constraints given as variations with respect to the steady state are:
if $-0.6\% < \Delta O_2 < -0.4\%$ with penalty weight: $w = 50$,
if $\Delta O_2 < -0.6\%$ with penalty weight: $w = 200$,
$|T_{\text{hot}}| < 20°\text{C}$ with penalty weight: $w = 3$.

Shell Control Problem

The Shell control problem (Prett et al., 1988) concerns a fractionator of heavy oil. The system possesses three manipulated variables (heat duty, side draws), seven outputs among which two controlled outputs (top and side draw compositions), four secondary outputs (temperatures which can be used for inferential control) and a seventh output to be maintained above a given value, and two unmeasured disturbances (top and intermediate reflux). This problem has been treated by several searchers among whom (Velez, 1997; Yu et al., 1994). The model and its uncertainties are given in (Prett et al., 1988). The control objectives are to maintain the controlled outputs at specifications, to minimize u_3 in order to maximize the heat recovery, to reject the disturbances, to obtain a certain closed-loop speed response, to respect constraints on the inputs, the input velocities, and on some outputs.

21.2.2 Description by a Linear State-Space Model

An Unstable Open-Loop System

Cheng (1989) considered the following discrete two-input two-output system

$$\begin{array}{rcl} \boldsymbol{x}(k+1) & = & \begin{bmatrix} -1.1 & 0.1 \\ 0.2 & -1.3 \end{bmatrix} \boldsymbol{x}(k) + \begin{bmatrix} 0.1 & 0.2 \\ 0 & 0.1 \end{bmatrix} \boldsymbol{u}(k) \\ \boldsymbol{y}(k) & = & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{x}(k) \end{array} \tag{21.24}$$

to test linear quadratic model algorithmic control. Note that this system is open-loop unstable.

Two-bed Exothermic Catalytic Reactor

Originally, this two-bed reactor was described by a set of nonlinear partial differential equations which have been later linearized and reduced. The model used (Foss et al., 1980) is this reduced linear state-space form.

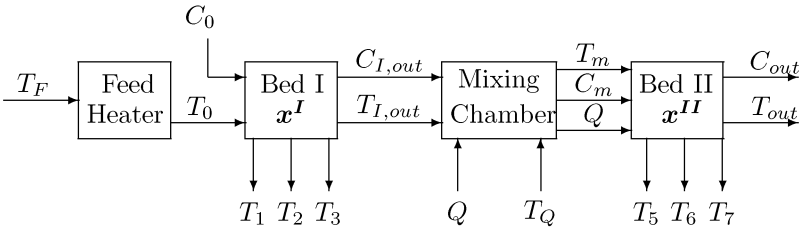


Figure 21.2: Decomposition of the process with inputs, internal variables and outputs

The decomposition of the process as shown in Fig. 21.2 allows us to understand the uncommon form of the linear state-space model
Feed heater:

$$\dot{T}_0 = -\frac{1}{\tau}\, T_0 + \frac{1}{\tau}\, T_f \tag{21.25}$$

Bed I:

$$\begin{aligned} \dot{\boldsymbol{x}}^I &= \boldsymbol{A}^I\, \boldsymbol{x}^I + \boldsymbol{B}^I \begin{bmatrix} T_0 \\ C_0 \end{bmatrix} \\ \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T^I_{out} \\ C^I_{out} \end{bmatrix} &= \boldsymbol{C}^I\, \boldsymbol{x}^I + \boldsymbol{D}^I \begin{bmatrix} T_0 \\ C_0 \end{bmatrix} \end{aligned} \tag{21.26}$$

Mixing chamber:

$$\begin{bmatrix} T_m \\ C_m \\ Q \end{bmatrix} = \boldsymbol{M} \begin{bmatrix} Q \\ T_Q \\ T^I_{out} \\ C^I_{out} \end{bmatrix} \tag{21.27}$$

Bed II:

$$\begin{aligned} \dot{\boldsymbol{x}}^{II} &= \boldsymbol{A}^{II}\, \boldsymbol{x}^{II} + \boldsymbol{B}^{II} \begin{bmatrix} T_m \\ C_m \\ Q \end{bmatrix} \\ \begin{bmatrix} T_5 \\ T_6 \\ T_7 \\ T_{out} \\ C_{out} \end{bmatrix} &= \boldsymbol{C}^{II}\, \boldsymbol{x}^{II} + \boldsymbol{D}^{II} \begin{bmatrix} T_m \\ C_m \\ Q \end{bmatrix} \end{aligned} \tag{21.28}$$

The data concerning the different matrices in the vicinity of the nominal steady state are given by Foss et al. (1980). The manipulated inputs are T_f, Q and T_Q. C_0 is a disturbance. The controlled outputs are T_{out} and C_{out}. $\boldsymbol{x}^I$ and $\boldsymbol{x}^{II}$ have dimension 7. In this article, the authors used the characteristic loci to design their compensators. They prefered that method to LQ control because of interpretation. It is clear that such a problem allows the comparison of many of the different methods commented on in this book.

21.2.3 Description by a State-Space Knowledge Model

A Semi-batch Reactor

Chin et al. (2000) use the model of a jacketed semi-batch reactor where exothermic series-parallel first-order reactions occur

$$\begin{aligned} &\mathrm{A} + \mathrm{B} \xrightarrow{k_1} \mathrm{C} && r_1 = k_1 C_A C_B \\ &\mathrm{B} + \mathrm{C} \xrightarrow{k_2} \mathrm{D} && r_2 = k_2 C_B C_C \end{aligned} \tag{21.29}$$

The reactor model is

$$\begin{aligned} \frac{dT}{dt} &= \frac{Q_{feed}}{V}(T_{feed} - T) - \frac{UA}{V\rho C_p}(T - T_j) - \frac{\Delta H_1}{\rho C_p} k_1 C_A C_B - \frac{\Delta H_2}{\rho C_p} k_2 C_B C_C \\ \frac{dC_A}{dt} &= -\frac{Q_{feed}}{V} C_A - k_1 C_A C_B \qquad C_A(t=0) = C_{A,0} \\ \frac{dC_B}{dt} &= \frac{Q_{feed}}{V}(C_{B,feed} - C_B) - k_1 C_A C_B - k_2 C_B C_C \\ \frac{dC_C}{dt} &= -\frac{Q_{feed}}{V} C_C + k_1 C_A C_B - k_2 C_B C_C \\ \frac{dV}{dt} &= Q_{feed} \end{aligned} \tag{21.30}$$

with the following initial conditions: $T(0) = T_0$, $C_A(t=0) = C_{A,0}$, $C_B(t=0) = 0$, $C_C(t=0) = 0$, $V(t=0) = V_0$ and the parameters defined in Table 21.6. The kinetic constants follow the Arrhenius law: $k_i = k_{i,0} \exp(-E_i/(RT))$.

Table 21.6: Initial variables and main parameters of the jacketed semi-batch reactor

Initial reactor volume	$V_0 = 50$ l
Feed flow rate	$Q_{feed} = \begin{cases} 0 & \text{if } t < 31\text{min} \\ Q_{feed}(t) & \text{if } t \geq 31 \end{cases}$
Feed temperature	$T_{feed} = 308$ K
Initial temperature	$T_0 = 298$ K
Initial concentration of reactant A	$C_{A,0} = 1$ mol.l^{-1}
Feed concentration of reactant B	$C_{B,feed} = 0.90$ or 0.95 mol.l^{-1}
Reduced heat transfer coefficient	$UA/(\rho C_p) = 0.375$ l.min^{-1}
Preexponential factor	$k_{10} = 5.0969 \times 10^{16}$ l.mol^{-1}.min^{-1}
Preexponential factor	$k_{20} = 2.2391 \times 10^{17}$ l.mol^{-1}.min^{-1}
Activation energy	$E_1/R = 12305$ K
Activation energy	$E_2/R = 13450$ K
Reduced heat of reaction	$\Delta H_1/(\rho C_p) = -28.5$ K.l.mol^{-1}
Reduced heat of reaction	$\Delta H_2/\rho C_p) = -20.5$ K.l.mol^{-1}
Sampling period	$T_s = 1$ min

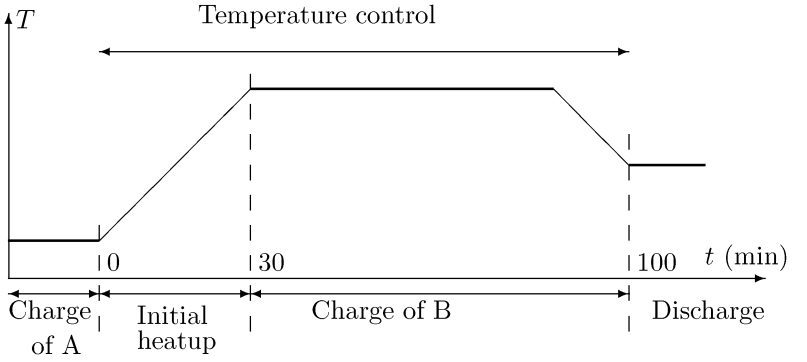


Figure 21.3: Temperature profile to be followed and operations for the semi-batch reactor

For such semi-batch reactors, a typical profile (Fig. 21.3) corresponds to the one given by Chin et al. (2000). The reactant B is loaded only between $t = 30$ min and the batch terminal time, which is fixed at $t = 100$ min. During the reaction period, it is assumed that the concentration of A is sampled every 10 min and measured with a delay of 5 min. The desired product is C and the objective is production of C equal to $V(t_f)C_C(t_f) = 42$ mol. The manipulated variables are the jacket temperature T_j and the flow rate Q_B of B. Several constraints are imposed

$$\begin{aligned} &298\ \text{K} \leq T_j(t) \leq 318\ \text{K} \\ &0.5\ \text{l.min}^{-1} \leq Q_B \leq 1.5\ \text{l.min}^{-1} \end{aligned} \tag{21.31}$$

This reactor model can be used for different studies: open-loop optimal control, model predictive control, observer estimation and batch reactor strategies.

A Cascade of two CSTRs

Gobin et al. (1994) describe a cascade of two polymerization CSTRs of industrial interest to which they applied model predictive control as DMC. This process is open-loop unstable and each reactor exhibits three steady states as in the polymerization reactor in Sect. 3.2.3. The reaction is the homopolymerization of styrene for which the kinetic scheme and data of Kim et al. (1990, 1991) are used. The temperature control is performed by manipulating the flow rates in the reactor jackets. By use of the polymerization reaction model, polymer properties such as average chain length and polydispersity can be calculated as well as the polymer concentration. Constraints are imposed on the coolant flow rates and their rate of change.

21.2.4 Continuous State-Space Models

Tennessee-Eastman Benchmark

The Tennessee-Eastman benchmark concerns a plant workshop which has been made available to the academic community to evaluate process monitoring, fault diagnosis and process control strategies. A detailed description is given by Downs and Vogel (1993). The plant produces two products and two by-products from four reactants. The process has five main unit operations: reactor, product condenser, vapour/liquid separator, recycle compressor and product stripper. The model of the process contains 50 differential equations, nonlinear and coupled. The open-loop simulation code for the process is written in Fortran and is available, for example, on the following web sites:
http://brahms.scs.uiuc.edu (accessed October 2003) or
http://depts.washington.edu/control/LARRY/TE/download.html (accessed October 2003) .

The process is open-loop unstable.

It has been used by many searchers (Chen, 1997; Duvall and Riggs, 2000; McAvoy and Ye, 1994; Ricker, 1995; Ricker and Lee, 1995a,b; Sriniwas and Arkun, 1997; Zheng, 1998)

Benchmark of Wastewater Treatment Plants

The control of wastewater treatment plants is difficult because of frequent and important changes of load in flow rate and in quality and also to the biological processes which are the fundamentals of the plant operation.

The International Water Quality Association and COST 624 group have established knowledge models representing the behaviour of wastewater treatment plants that can be used to test estimation, diagnostic and control strategies. COST 624 group published a benchmark (Alex et al., 2002; Carlsson and Rehnström, 2002; Pons et al., 1999; Vrecko et al., 2002), also available on the following web site:
www.ensic.inpl-nancy.fr/COSTWWTP/ (accessed October 2003)

The simulated wastewater treatment plants possesses a series of five reactors, the first two ones being mixed and non-aerated, the three following ones simply aerated; this group is followed by a secondary settler. Two recycle streams complete the process. The model of the biological process is ASM1 of IAWQ and includes 13 components and eight reaction processes. Typical feed disturbances for dry, stormy or rainy weather are available as representative files of 14 days with a sampling period of 15 min. Performance criteria have been established concerning the effluent quality; constraints corresponding to the operating norms are imposed on the effluents and operating costs are proposed.

References

A. Al-Ghazzawi, E. Ali, A. Nouh, and E. Zafiriou. On-line tuning strategy for model predictive controllers. *J. Proc. Cont.*, 11:265–284, 2001.

J. Alex, J.F. Béteau, J.B. Copp, J. Dudley, R. Dupont, S. Gillot, U. Jeppsson, J.M. LeLann, M.N. Pons, and P. A. Vanrolleghem, editors. *The COST Simulation Benchmark. Description and Simulator Benchmark*. European Communities, 2002. ISBN 92-894-1658-0.

G. Birol, C. Undey, and A. Cinar. A modular simulation package for fed-batch fermentation: penicillin production. *Comp. Chem. Engng.*, 26:1553–1565, 2002.

B. Carlsson and A. Rehnström. Control of an activated sludge process with nitrogen removal - a benchmark study. *Water Science and Technology*, 45 (4–5):137–142, 2002.

H. Chen. *Stability and Robustness Considerations in Nonlinear Model Predictive Control*. PhD thesis, Stuttgart University, 1997.

C.M. Cheng. Linear quadratic-model algorithmic control method: a controller design method combining the linear quadratic method and the model algorithmic control algorithm. *Ind. Eng. Chem. Res.*, 28:178–186, 1989.

Y. Chikkula and J.H. Lee. Robust adaptive predictive control of nonlinear processes using input-output models. *Ind. Eng. Chem. Res.*, 39:2010–2023, 2000.

I.S. Chin, K.S. Lee, and J.H. Lee. A technique for integrated quality control, profile control and constraint handling for batch processes. *Ind. eng. Chem. Res.*, 39:693–705, 2000.

J. Choi, H.S. Ko, and K.S. Lee. Constrained linear quadratic optimal control of chemical processes. *Comp. Chem. Engng.*, 24:823–827, 2000.

C.R. Cutler and B.L. Ramaker. Dynamic matrix control - a computer control algorithm. In *Joint Automatic Control Conference*, number WP5-B, San Francisco, CA, 1980.

P.B. Deshpande. *Multivariable Process Control*. Instrument Society of America, North Carolina, 1989.

J.J. Downs and E.F. Vogel. A plant-wide industrial process control problem. *Comp. Chem. Engng.*, 17(3):245–255, 1993.

P.M. Duvall and J.B. Riggs. On-line optimization of the Tennessee Eastman challenge problem. *J. Proc. Cont.*, 10:19–33, 2000.

A.S. Foss, J.M. Edmunds, and B. Kouvaritakis. Multivariable control system for two-bed reactors by the characteristic locus method. *Ind. Eng. Chem. Fundam.*, 19:109–117, 1980.

F. Gobin, L.C. Zullo, and J.P. Calvet. Model predictive control of an open-loop unstable train of polymerization reactors. *Comp. Chem. Engng.*, 18: s525–s528, 1994.

M.A. Henson and D.E. Seborg. Theoretical analysis of unconstrained nonlinear model predictive control. *Int. J. Cont.*, 58(5):1053–1080, 1993.

K.J. Kim, K.Y Choi, and J.C. Alexander. Dynamics of a cstr for styrene polymerization initiated by a binary initiator system. *Polym. Eng. Sci.*, 30 (5):279–290, 1990.

K.J. Kim, K.Y. Choi, and J.C. Alexander. Dynamics of a cascade of two continuous stirred tank styrene polymerization reactors with a binary initiator system. *Polym. Eng. Sci.*, 31(5):333–352, 1991.

P. Lunström, J.H. Lee, M. Morari, and S. Skogestad. Limitations of dynamic matrix control. *Comp. Chem. Engng.*, 19(4):409–421, 1995.

T.J. McAvoy and N. Ye. Base control for the Tennessee Eastman challenge problem. *Comp. Chem. Engng.*, 18:383–413, 1994.

M.N. Pons, H. Spanjers, and U. Jeppsson. Towards a benchmark for evaluating control strategies in wastewater treatment plants by simulation. *Comp. Chem. Engng.*, 23S:403–406, 1999.

D.M. Prett and C.E. Garcia. *Fundamental Process Control.* Butterworths, Stoneham, MA, 1988.

D.M. Prett, C.E. Garcia, and B.L. Ramaker. *The Second Shell Process Control Workshop.* Butterworths, Stoneham, MA, 1988.

W.F. Ramirez. *Process Control and Identification.* Academic press, New York, 1994.

N.L. Ricker. Optimal steady-state operation of the Tennessee Eastman challenge process. *Comp. Chem. Engng.*, 19(9):949–959, 1995.

N.L. Ricker and J.H. Lee. Nonlinear model predictive control of the Tennessee Eastman challenge process. *Comp. Chem. Engng.*, 19(9):961–981, 1995a.

N.L. Ricker and J.H. Lee. Nonlinear modeling and state estimation for the Tennessee Eastman challenge process. *Comp. Chem. Engng.*, 19(9):983–1005, 1995b.

J.A.D. Rodrigues and M. Filho. Optimal feed rates strategies with operating constraints for the penicillin production process. *Chem. Eng. Sci.*, 51:2859–2864, 1996.

B. Srinivasan, D. Bonvin, E. Visser, and S. Palanki. Dynamic optimization of batch processes II Role of measurements in handling uncertainty. *Comp. Chem. Engng.*, 27:27–44, 2002a.

B. Srinivasan, S. Palanki, and D. Bonvin. Dynamic optimization of batch processes I characterization of the nominal solution. *Comp. Chem. Engng.*, 27:1–26, 2002b.

G.R. Sriniwas and Y. Arkun. Control of the Tennessee Eastman process using input-output models. *J. Proc. Cont.*, 7:387–400, 1997.

A.J. Stack and F.J. Doyle III. Application of a control-law nonlinearity measure to chemical reactor analysis. *AIChE J.*, 43:425–447, 1997.

A. Uppal, W.H. Ray, and A.B. Poore. On the dynamic behaviour of continuous stirred tank reactors. *Chem. Eng. Sci.*, 29:967–985, 1974.

J.C. Velez. Linear Programming Based Model Predictive Control for the Shell Control Problem. Master's thesis, Purdue University, 1997.

D. Vrecko, N. Hvala, and J. Kocijan. Wastewater treatment benchmark: what can be achieved with simple control? *Water Science and Technology*, 45 (4–5):127–134, 2002.

P. Vuthandam, H. Genceli, and M. Nikolaou. Performance bounds for robust quadratic dynamic matrix control with end condition. *AIChE J.*, 41(9): 2083–2097, 1995.

Z.H. Yu, W. Li, J.H. Lee, and M. Morari. State estimation based model predictive control applied to Shell control problem: a case study. *Chem. Eng. Sci.*, 49(3):285–301, 1994.

E. Zafiriou. An operator control theory approach to the Shell standard control problem. In *Shell Process Control Workshop*, Stoneham, MA, 1990. Butterworths.

R. Zanovello and H. Budman. Model predictive control with soft constraints with application to lime kiln control. *Comp. Chem. Engng.*, 23:791–806, 1999.

A. Zheng. Nonlinear model predictive control of the Tennessee Eastman process. In *American Control Conference*, Philadelphia, 1998.

Index

Action
  derivative, 93
  integral, 90
  proportional, 86
Actuators, 71
Adaptation
  gain matrix, 426
Adjoint variable, 509
Aliasing, 333
Analog scheme, 41
Anti-aliasing filter, 334
Anti-windup, 150
AR, 392
ARARMAX, 393
ARARX, 393
ARIMA, 392
ARMA, 392
ARMAX, 390
ARX, 388
Asymptotic stability domain, 106
Attractor, 106
Augmented function, 500
Autocorrelation, 317, 324
Autocovariance, 324
Autonomous, 104

Balance
  energy, 15
  mass, 10
Bang-bang, 517, 518
Bayes theorem, 413
Bellman, 525
Bezout equation, 167, 170, 466, 558
Bilinear integration, 362
Black plot, 195
Bode
  plot, 188
  stability criterion, 204
Boiler, 249
Bolza, 497
Boxcar, 329
Brunovsky canonical form, 636
Byrnes-Isidori canonical form, 639

Canonical form of Jordan, 266
Cascade, 244
Causality, 361
Certainty equivalence principle, 490
Characteristic
  equation, 119
  loci, 283
  polynomial, 40, 266, 270
Chemical reactor
  generalized predictive control with performance model, 566
  generalized predictive control, 561
  identification, 456
  internal model control, 486
  isothermal, 12, 30
  non-isothermal, 14
  nonlinear geometric control, 682
  pole-placement control, 477
Cohen-Coon, method of, 143
Complex numbers, 185
Condition number, 266, 297, 422, 577
Conditions, terminal, 500
Control, 5
  adaptive, 488
  by ratio, 258
  composition, 136
  discrete internal model, 483
  dual, 714
  flow rate, 135
  gas pressure, 136
  generalized predictive, 555
  level, 135
  linear quadratic, 175
  model predictive, 575
  model reference, 463
  nonlinear geometric, 619
  optimal, 493
  pole-placement, 463

selective, 249
split-range, 250
temperature, 136
variable, 66
Controllability, 264
Controllability, integral, 296
Controllable
canonical form, 265
companion form, 265
Controller, 67
digital PID, 481
feedforward, 251
one degree of freedom, 174
proportional, 67
proportional-integral, 67
proportional-integral-derivative, 68
RST, 463
self-tuning, 719
two degrees of freedom, 166
Convolution product, 26
Corrector, 67
Correlation analysis, 374
Costate, 509, 513
Covariance matrix, 660
Cramer-Rao inequality, 414
Criterion
of Nyquist, generalized inverse, 284
Bode stability, 204
IAE, 132
ISE, 131
ITAE, 132
of Jury, 359
of Nyquist, 213
of Nyquist, generalized, 284
of Routh-Hurwitz, 119
Cross-correlation, 317, 325
Cross-covariance, 325
Crystallizer, 153

Damping, 46
Decibel, 189
Decoupling, 286, 645
Degree (one) of freedom controller, 174
Degree of freedom, 20
Degrees (two) of freedom controller, 166
Delay, 25, 53
Delay margin, 225
Derivative, Lie, 628
Design of Feedback Controllers, 129
Diffeomorphism, 632
Differential geometry, 627
Diophantine equation, 167, 170, 558
Dirac, 6, 43
Direct Synthesis Method, 159
Discontinuity condition, 501
Discrete internal model control, 483
Discrete transfer function, 349
Distillation, 250, 252, 258, 522, 705
column of Wood and Berry, 287
composition estimator, 659
extractive, 536
extractive column, linear quadratic control, 536
extractive, linear quadratic Gaussian control, 544
NMPC, 600
robustness study, 303
Distributed-parameter, 18
Distribution, 631
integrable, 631
involutive, 631
Disturbance, 5, 66
model, 445
DMC tuning, 589
DMC, dynamic matrix control, 581
Dominant pole, 131
Dynamic programming, 525
Dynamic state feedback, 635, 653
Dynamics
of zero, 624
forced, 624
internal, 624
unforced, 624

Eigenvalue, 40, 105, 275
Energy, 321
Balance, 15
Equation
Bezout, 466
characteristic, 119
error model, 389
Hamilton-Jacobi, 502, 509
Hamilton-Jacobi-Bellman, 529
Lyapunov, 109
minimal, 271
Ergodic, signal, 324
Error, 131
equation, 389
of acceleration, 133

of position, 133
of velocity, 133
prediction a priori, 385
Euler
conditions, 499, 500, 506
method, 362
Euler-Lagrange lemma, 497
Evans locus, 121
Evaporator, 652
Exogenous, 389
Expectation, mathematical, 323
Extended Kalman filter, 664, 687

Fast Fourier transform, 321
FCC, 600
Feedback, 65, 82
Feedback difference, 280
Feedforward controller, 251
Filter, discrete Kalman, 398
First-order, 44
Fisher information matrix, 414
Fluid catalytic cracking, 600
Focus, 106
Forced
dynamics, 624
response, 357
Form
canonical of Brunovsky, 636
canonical of Byrnes-Isidori, 639
canonical of Jordan, 266
controllable canonical, 265
controllable companion, 265
modal canonical, 266
normal, 623, 633
observable canonical, 269
observable companion, 269
of Smith-McMillan, 282
Fourier
discrete transform, 320
inverse discrete transform, 320
series expansion, 318
transform, 314
Frequency
analysis, 183
crossover, 141
gain crossover, 208
phase crossover, 201
specification, 229
warping, 363
Frobenius theorem, 630

Function
Lyapunov, 107
of repartition, 323
of sensitivity, 223
of sensitivity, complementary, 223
orthogonal, 318
Functional, 495

Gain, 44
crossover frequency, 208
margin, 207, 225
ultimate, 136
Generalized predictive control, 555
Generic model control, 647
Gershgorin circles, 284, 298
Gibbs phenomenon, 320
Globally linearizing control, 646
Globally linearizing observer, 666
GPC, generalized predictive control, 555
Gradient, 423
Grammian
of controllability, 272
of observability, 272

Hamilton conditions, 503, 509
Hamilton-Jacobi, 502, 509
Hamilton-Jacobi equation, 503
Hamilton-Jacobi-Bellman, 529
Hamiltonian, 502
High-gain observer, 667
Holder, 329
Horizon
of control, 583
of model, 582
of prediction, 584
Hurwitz, 120

IAE criterion, 132
Identification, 57
state-space, 397
Akaike criterion, 452
IMC, internal model control, 160
Impulse, 43
Innovation, 385
Innovation form, 403
Input, 5
Input-output linearization, 638
Input-state linearization, 635
Instrumental variable, 416, 443
Integral curve, 104
Integral saturation, 150

Integrator, 45, 147
Integrity, 296
Interaction, 286
Internal
    dynamics, 624
    model control, 160
    model control, discrete multivariable, 307
    model principle, 220
    stability, 640
Inverse response, 55, 241
Involutivity, 631
ISE criterion, 131
ITAE criterion, 132

Jacobian matrix, 497, 629
Joint probability, 323
Joint probability density, 323
Jordan canonical form, 266
Jury, stability criterion, 359

Kalman
    discrete filter, 398
    gain matrix, 399
Kalman filter, 593
    continuous-discrete, 664
    discrete, implementation, 405
    discrete-discrete, 664
    estimator, 401
    extended, 664
    predictor, 401

Lagrange, 497
Laplace, 21
Laurent series, 343
Least Squares, 409
Least squares, 421
    recursive, 431
    recursive extended, 438
    recursive generalized, 438
    simple recursive, 429
Legendre-Clebsch condition, 502
Lemma matrix inversion, 400
Lie
    bracket, 629
    derivative, 628
Likelihood function, 413
Limit cycle, 107
Linear programming, 595
Linear quadratic control, 530
    by transfer function, 175
Linear quadratic Gaussian control, 541
Linearization, 9, 22
    input-output, 638
    input-state, 635
Luenberger, observer, 663
Lumped-parameter, 10
Lyapunov
    direct method, 107
    equation, 109, 272
    function, 107
    stability, 105

MA, 392
Margin
    delay, 225
    gain, 207, 225
    modulus, 228
    phase, 208, 225
    stability, 226
Markov
    parameters, 263, 621
    series, 263
Mason formula, 77
Mass balance, 10
Matrix
    inversion lemma, 400
    Jacobian, 629
    norm, 297
    normal, 302
    of transfer functions, 279
    orthogonal, 291
    orthonormal, 291, 422
    state transition, 262, 398
    Sylvester, 171
    symplectic, 550
    trace, 400
Maximum likelihood, 412
Maximum likelihood recursive, 439
Maximum principle, 511
Mayer, 497
Mean, 323
MIMO, 6
Minimum principle, 511
Minimum-phase, 104, 190, 627, 643
Modal canonical form, 266
Mode, 35
Model
    AR, 392
    ARARMAX, 393
    ARARX, 393

ARIMA, 392
ARMA, 392
ARMAX, 390
ARX, 388
Box-Jenkins, 395
dynamic, 7
equation error, 388, 389
general, 395
MA, 392
output error, 393
reference, 469
RIF, 393
steady-state, 7
Model predictive control, 575
Modulus margin, 228
Moment, 60, 323
Moment method, 60
Monic, polynomial, 168
Monitoring, 659
Moving average, 339
Moving state estimation, 669
MPC, model predictive control, 575
Multivariable, 6
control, 279
nonlinear control, 647

Natural response, 357
Nichols plot, 217
Niederlinski index, 286
NIPALS, 661
Node, 106
Noise, white, 373
Noise-spike, 339
Nominal performance, 224
Nonlinear multivariable control, 647
Nonminimum-phase, 190–192
Norm
of a signal, 222
of a transfer function, 222
of matrix, 297
Normal form, 623, 633
Normalized time, 451
Nyquist
criterion, 213
frequency, 363
generalized criterion of, 284
generalized inverse criterion of, 284
plot, 192
point, 214

Observability, 268
Observable
canonical form, 269
companion form, 269
Observation
vector, 389
Observer, 657
globally linearizing , 666
high-gain, 667
Luenberger, 663
moving state estimation, 669
polynomial, 471
On-Off control, 151
Operator
δ, 366
backward shift q^{-1}, 365
forward shift q, 365
Optimal control, 493
Optimality principle, 525
Optimization, 57
Order, 20
Orthogonal
function, 318
matrix, 291
Orthonormal matrix, 291, 422
Output, 5, 66
Output error method, 444
Overshoot, 49

Padé approximation, 54
Pairing, 286
Parameter
adaptation algorithm, 424
distributed, 7, 18
lumped, 7, 10
vector, 389
Parametric identification, models, 387
Parseval-Plancherel, 27, 316
Partial least squares, PLS, 659
Partial state, 42
PCA, principal component analysis, 659
Performance
index, 497
nominal, 224
robust, 227
pH control, 154
Phase
crossover frequency, 201
margin, 208, 225
plane, 106
portrait, 106

Physical realizability, 118
PID controller tuning, 136
Plot
  Black, 195
  Bode, 188
  Nichols, 217
  Nyquist, 192
PLS, partial least squares, 659
Point
  equilibrium, 105
  singular, 105
  stationary, 105
Pole, 35, 357
  dominant, 131
  multivariable, 282
  placement, 166, 463
Pole-zero correspondence, 363
Polymerization reactor, 111, 524, 653, 654
Pontryagin maximum principle, 511
Prediction
  error, a posteriori, 424
  error, a priori, 424
Predictor
  a posteriori, 424
  a priori, 424
  multi-step, 385
  one-step, 380
Principal component analysis, PCA, 659
Principal direction, 292
Principle
  of internal model, 220
  of superposition, 33
Probability density, 323
Process, staged, 17
Projection to latent structures, PLS, 659
Proper transfer function, 361
Pseudo random binary sequence, 448

QDMC, quadratic dynamic matrix control, 589
Quantization, 313

Random signal, 322
Random stationary signal, 324
Rank, 265
Rate of reaction, 13
Reachability, 264
Reactor
  biological, nonlinear control, 692
  catalytic, 247, 250
  chemical, 250
  chemical tubular, 18
  chemical, generalized predictive control with performance model, 566
  chemical, generalized predictive control, 561
  chemical, internal model control, 486
  chemical, isothermal, 12
  chemical, non-isothermal, 14
  chemical, nonlinear control, 681
  chemical, pole-placement control, 477
  identification, 456
  polymerization, 111, 653, 654
  with jacket, 244
Realization, 262, 271
  balanced, 273
  minimal, 272
Reconstruction, 341
Recursive
  least squares, 431
  maximum likelihood, 439
  prediction error, 440
Reduction of model, 273
Reference, 148
  model, 469
  trajectory, 148
Regression, pseudo-linear, 392
Regulation, 65, 88
Relative
  degree, 122, 125, 361, 621, 629, 648
  gain array, 292
  order, 621
Relay oscillation method, 138
Response
  forced, 43, 357, 559
  free, 559
  impulse, 43
  inverse, 55, 241
  natural, 43, 357
  step, 43
RGA, 292
Riccati
  discrete equation, 401
  equation, 533, 541, 542

RIF, 393
Robust
performance, 227
stability, 226
Robustification, 444
Robustness, 173, 221, 296, 298
Root locus, 121
Routh-Hurwitz, 119

Saddle point, 106
Sampling, 329
Sampling period, 335
Second-order, 46
Sensitivity
complementary function, 299
complementary function of, 223
function, 223, 299
Sensor, 69
intelligent, soft, 657
Separation principle, 543
Series, Fourier, 318
Set point, 66
Shannon, 334
Signal
ergodic, 324
random, 322
random stationary, 324
stochastic, 322
stochastic stationary, 324
Signal-flow graph, 77
Single variable, 6
Singular
arc, 514, 520
value, 291
value decomposition, 291
Singular value decomposition, 661
SISO, 6
Smith predictor, 239
Smith-McMillan, form, 282
Smoothing filter, 339
Soft
sensor, 657
Specification frequency, 229
Spectral
density, 317, 324
factorization, 175
Spectrum, 317
Square system, 280
Stability, 20, 103, 282, 372
asymptotic, 105
asymptotic domain, 106
external, 103
internal, 103, 640
Jury criterion, 359
Lyapunov, 105
margin, 226
polymerization reactor, 111
robust, 226
state-space, 104
strict, 372
Stabilizability, integral, 295
State
estimator, 657
representation, 42, 261
space, 39
transition matrix, 398
State feedback
dynamic, 635
static, 635
State-space, 8
Static state feedback, 635
Stationary, signal, 324
Statistical process control, 659
Step, 43
Stochastic signal, 322
Stochastic stationary signal, 324
Superposition principle, 33
Surge tank, 10, 30

Theorem, Cauchy, 213
Time
constant, 44
delay, 53
rising, 49
settling, 49
Total finite energy, 316
Trace of matrix, 400
Tracking, 65, 87
Trajectory, 104
Transfer function
continuous, 27
discrete, 349
proper, 29
Transform
z, 343
fast Fourier, 321
Fourier inverse discrete, 320
Laplace, 21
of Fourier, 314
Transversality conditions, 500

Tuning of PID, 136
Tustin method, 362

Ultimate
    gain, 136, 203
    period, 137
Uncertainty relation, 316
Unforced dynamics, 624

Validation, 447
Variable
    control, 65
    controlled, 65
    deviation, 22
    manipulated, 5
Variance, 323
Variational method, 496
Volterra model, 597

Warping, 363
Weierstrass-Erdmann
    condition, 502
White noise, 373
Windup, 150

z
    -transform, 343, 347
Zero, 35, 357
    dynamics, 624, 639
    multivariable, 282
Zero-order holder, 346
Ziegler-Nichols tuning, 137